Detection of frequency and intensity changes using synthetic vowels and other sounds

Paul Cosgrove

A thesis submitted to the University of Keele in partial fulfilment of the requirements for the Degree of Doctor of Philosophy

September 1988

Wagner pre-empting psychoacoustical methodology

ABSTRACT

Formant frequency transition detection thresholds in synthetic vowels were investigated for their dependence on formant number, vowel type and duration of frequency transition. With transitions in final position it was found that F2 is easier to detect than F1 for all tested vowels when thresholds are expressed in terms of critical bands.

The proximity of neighbouring formants appeared to affect the thresholds; lower values were obtained for formants that were positioned in excess of a critical bandwidth from neighbouring formants. All thresholds exhibited a decreasing tendency with increasing transition duration.

An excitation pattern model was used to compare thresholds from each of the experimental conditions. This was effective in normalising the data and confirming a masking hypothesis as an explanation for threshold differences.

Thresholds were also obtained for sinewave stimuli. These, as one would expect, proved to be superior to those obtained for formants (on average by 4.5:1) though the ratio was dependent on frequency.

Comparisons were made between frequency transitions for pure tones and DL data, a well-documented area of psychoacoustics. The data show a similar relationship, although discrimination thresholds were consistently lower than transition thresholds.

Frequency transition detection thresholds in initial position were investigated for their alleged inferiority compared with final position transitions. This was confirmed universally by experiments on both formant and sinewave transitions. For the vowel stimuli inconsistencies with a masking model were found. This suggested the

possibility of a speech mode of perception that operated predominantly when speech cues were more powerful.

Finally, intensity transition detection thresholds (and difference limens) were obtained for synthetic vowel stimuli. Experimental conditions were analogous to those used for frequency transition detection. Similar functions were obtained, and vowel type proved to be unimportant. However, transitions of increasing intensity proved to be easier to detect than those of decreasing intensity.

Acknowledgements

There are a number of people that I wish to thank for their help towards the execution and completion of this project:-

Pat Wilson for his supervision and perseverance. Roger Moore and the Ministry of Defence for financial support of the project. John Ruscoe for able technical assistance. Dave Parker for some excellent graph plotting software, and having the patience to show me how to use it. Roddy Pratt and David Foster for statistical advice. Bill Ainsworth and Ted Evans for answering many questions about speech and hearing.

The modelling of the data was achieved with the kind assistance of Roy Patterson and John Holdsworth. Roy's software filterbank and John's 'pipes' were invaluable in making some sort of sense out of a mass of confused looking data.

It would not have been possible to finish the thesis after leaving Keele without the selfless help of Margaret Hodgson in the department and Steve Morris in the Computer Centre.

I also wish to thank all the subjects who took part in the experiments, even those who had to be dragged kicking and screaming to the sound-proofed booth.

I would especially like to thank my parents whose interest, encouragement and support throughout my years in education has been as much as anyone could wish for.

Finally, I would like to thank Debbie for not only typing a substantial part of the text and completing many hours of experiments, but mainly for encouraging me to continue when all I wanted to do was give up.

A special mention must also be made of the West Midlands Fire Service who, in their enthusiasm to save parts of my writing from a burning canal boat, sank it. This was almost one setback too many. Of course, I shall always maintain that the best of the thesis went down with the ship.

Needless to say, all the faults are mine.

Glossary of Abbreviations

A	amplitude
AM	amplitude modulation
CB	critical bandwidth (Hz)
CF	characteristic frequency
CFN	comb-filtered noise
cps	cycles per second
CV	consonant-vowel
dB	decibel(s)
DL	difference limen
ERB	equivalent rectangular bandwidth
F, f	frequency (Hz)
ΔF, Δf	frequency increment (Hz)
F0	fundamental frequency
F1, F2 etc	first, second formant etc
Fig., Figs	figure(s)
FM	frequency modulation
FTC	frequency threshold curve
Hz	Hertz (cycles per second)
I	intensity (normally, but is also used to mean amplitude, level etc.)
ΔI	intensity increment (dB)
JND	just noticeable difference
L	level (dB)
ms	milliseconds
PTC	psychophysical tuning curve
s	seconds
SL	sensation level (dB)
SPL	sound pressure level (dB)
VOT	voice onset time
2AFC	two-alternative forced-choice
2IFC	two-interval forced-choice

Representation of Phonetic Symbols

Some of the phonetic symbols used in the following text are not standard representations as published in the International Phonetic Alphabet. Where it was impossible to represent a phonetic symbol with a literal typed transcription an alternative symbol was chosen from the ASCII set of codes. The 'mapping' of phonetic symbols onto ASCII characters is detailed in Wells (1987), and the choices of alternative representations are based on his recommendations.

CONTENTS

Page No.

Chapter 1 General Introduction and Literature Survey..... 1

1.1 General Introduction..... 1
1.2 Frequency Resolution and Masking..... 4
1.2.1 Introduction..... 4
1.2.2 Masking..... 4
1.2.3 Critical Bands and Critical Ratios..... 6
1.2.4 Loudness Summation..... 7
1.2.5 Auditory Filter Shape..... 8
1.2.6 The Psychophysical Tuning Curve..... 12
1.2.7 Lateral Suppression..... 12
1.2.8 Conclusions..... 13
1.3 Frequency Discrimination..... 14
1.3.1 Introduction..... 14
1.3.2 Theories of Pitch Perception..... 14
1.3.2.1 Place Models..... 14
1.3.2.2 Temporal Models..... 15
1.3.3 Experiments to Determine Frequency DLs..... 16
1.3.4 Frequency Discrimination and Speech Perception 19
1.4 Frequency Modulation..... 23
1.4.1 Introduction..... 23
1.4.2 Frequency Transition Detection and Pitch Perception..... 24
1.4.3 Differences in Sensitivity to Upward and Downward Sweeps..... 27
1.4.4 Adaptation Studies with FM Stimuli..... 28
1.4.5 The Role of Formant Transitions in Speech Perception..... 29
1.4.6 Frequency Discrimination and Frequency Transition Detection: Differences and Similarities..... 32
1.4.7 Conclusions..... 34
1.5 Intensity Discrimination..... 35
1.5.1 Introduction..... 35
1.5.2 Measurement of Intensity DLs..... 35
1.5.3 Intensity Coding..... 38
1.5.4 Intensity Discrimination and Speech Perception 44
1.6 Evidence For and Against a Speech Mode of Perception..... 48

Chapter 2 Methods 51
2.1 Generation of Stimuli..... 51
2.2 Apparatus and Procedure..... 55
2.2.1 Apparatus..... 55
2.2.2 Procedure..... 56
2.3 Subjects..... 59
2.4 Analysis of Data..... 59

Chapter 3 Frequency Transition Detection: A Pilot Study..... 61
3.1 Abstract..... 61
3.2 Methods..... 61
3.3 Results..... 63
3.4 Discussion..... 66

Chapter 4 Frequency Transition Detection Using Synthetic Vowel Sounds..... 73
4.1 Introduction..... 73
4.2 Methods..... 73
4.2.1 Choice of Stimuli..... 73
4.2.2 Other Stimulus Parameters..... 74
4.3 Analysis of Data..... 75
4.4 Results 1..... 76
4.5 Discussion 1..... 80
4.5.1 Temporal Integration..... 80
4.5.2 Arguments For Representing Thresholds on a Bark Scale..... 84
4.6 Results 2..... 86
4.6.1 Contrast Analysis of Threshold Pairs Expressed in Terms of Critical Bands..... 86
4.6.2 Contrast Analysis of Threshold Pairs with the 10 ms Data Omitted..... 89
4.7 An Excitation Pattern Model of the Detection of Formant Frequency Transitions..... 91
4.7.1 Exposition of the Model..... 91
4.7.2 Testing the Model..... 95
4.8 Discussion 2 and Conclusion..... 102

Chapter 5 Frequency Transition Detection Using Pure Tones .. 104
5.1 Introduction..... 104
5.2 Methods..... 105
5.3 Results..... 106
5.4 Contrast Analysis of Threshold Pairs with the 10 ms Data Omitted..... 111
5.5 Comparison of Results..... 111
5.6 Summary and Conclusions..... 116

Chapter 6 Frequency Transition Detection With Transitions In Initial Position..... 118
6.1 Introduction..... 118
6.2 Methods..... 118
6.3 Results and Discussion..... 120
6.4 Discussion 2 and Conclusions..... 125

Chapter 7 Intensity Transition Detection and Discrimination. 129
7.1 Introduction..... 129
7.2 Methods..... 129
7.3 Results and Discussion..... 131
7.4 Conclusions..... 134

Chapter 8 Summary and Conclusions..... 135
8.1 Summary..... 135
8.2 Caveats and Criticisms..... 137
8.3 Recommendations for Further Studies..... 139
8.4 Final Comments..... 140

Appendix A Instructions to Subjects..... 142

Appendix B Methodology of Statistical Testing..... 144

Appendix C Supplementary Tables..... 147

References 176

CHAPTER 1

General Introduction and Literature Survey

1.1 General Introduction

The methods used to attempt the solution of the problem of machine recognition of speech have changed radically over the past few years. Previous approaches have been dominated by the use of whole word templates, matching patterns of single words in a spoken sentence with a stored vocabulary. Dynamic time-warping is an algorithm which employs the distortion of a spoken word (in time) in order to make a best fit to one word from a set in a stored vocabulary. The means by which this is achieved is by dynamic programming, a method almost universally used in automatic speech recognition up until the early 1980's (Moore, Beardsley et al, 1982; Moore et al, 1982).

Kay (1987) in a review of the UK Alvey Research Programme (in speech technology) includes much criticism of this "mechanistic" approach. The argument is that template matching is unsatisfactory as a general paradigm as it is "...only practical for limited task domains through the availability of powerful processing and the now long standing familiarity with the relevant algorithms."

Peckham (1986) argues the case slightly differently, stating that "...such machines are inherently limited in their ability to cope with variation in speech and the effects of coarticulation. A spoken word in isolation can differ dramatically from the same spoken word in a phrase or sentence." (Coarticulation is the movement of several articulators (of the speech production system) at the same time in anticipation of future utterances).

Such realisations of the drawbacks of previous approaches have led many workers to consider modelling the auditory system which, after all, is the most error free speech recognition system available. A large amount of contemporary work is concerned with using current knowledge of auditory psychophysics and physiology (Meddis, 1986) and from that attempting, amongst other things, "...to establish a central 'theory' of speech pattern processing." (Moore and Bridle, 1986). It is with such aims in mind that the current work was commissioned and undertaken.

Psychoacoustic knowledge in the areas of frequency discrimination, resolution and modulation, and also pitch perception and intensity discrimination is quite detailed. There is a great need for systematic psychoacoustic studies using stimuli that resemble, and are perceived as, speech. The stimuli used for investigations into the above subject areas, however, have tended to be simple in nature (e.g. pure tones, noise etc). The main aim of the current study has been to "...reinvestigate them under conditions which are more representative of those occurring in running speech." (Wilson, 1983). So by using synthetic speech stimuli, a number of auditory abilities have been tested for their qualitative and quantitative differences to less spectrally complex sounds (whose measurements are already well-documented).

The findings of psychoacoustic studies, such as the present one, have applications in providing more realistic parameters for preprocessing prior to the extraction of features/patterns from a speech signal.

Before the literature concerning frequency transition detection and intensity discrimination can be reviewed some basic psychophysical concepts must first be discussed. In order that the

factors involved in detecting a transition in frequency can be understood it is necessary to deal with how the auditory system resolves a sound into separate frequency components; the concepts of masking and perception of pitch must also be discussed. The following literature survey is organised in terms of introducing such concepts in a logical sequence rather than in order of importance.

1.2 Frequency Resolution and Masking

1.2.1 Introduction

Frequency resolution demonstrates the ability to separate two or more tones of different frequency that are simultaneously presented. The basilar membrane in combination with active mechanical feedback is the initial and major stage of this process; though it is thought that there may be some involvement of higher centres.

Most workers assume, with some justification, that the peripheral auditory system behaves like a set of overlapping linear band-pass filters. Linear in this case refers to the invariance of bandwidth over level. For a complex signal to be fully resolved by the auditory system, each component of the signal must be passed down a different filter channel.

It is important to ascertain the characteristics of auditory filters to fully understand the concept of frequency resolution. The two most important features of filters are centre frequency and bandwidth. A secondary factor, important in understanding masking, is the shape of the auditory filter.

1.2.2 Masking

Virtually all experiments to derive auditory filter shape and bandwidth use masking techniques. Masking has been defined as:

(1) The process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound.

(2) The amount by which the threshold of audibility of a sound is raised by the presence of another (masking) sound. The unit customarily used is the decibel.

(American Standards Association, 1960).

Although this definition is not ideal in some respects (Moore, 1982 points out the problem of confounding the observed effect with the physical process) it should partially suffice to explain the phenomenon.

Masking is currently thought to be caused by either the swamping of activity evoked by the signal, or the suppression of activity that the signal would evoke if it were presented in isolation. It is not known which of these mechanisms accurately accounts for the phenomenon of masking, though there are good arguments for both. Another possible explanation of the mechanism of masking lies in what is known as the 'energy detector' model, (e.g. Jeffress, 1970; Patterson, 1976; Pick, 1980). Essentially, the model explains that masking takes place if the stimulus does not cause a just-noticeable increase in the intensity of a stimulus plus masker signal. (The term intensity refers to the energy or power contained in a signal).

There are quite a few different types of masking, and these can be divided into two groups, simultaneous and non-simultaneous. Of the non-simultaneous types, forward and backward masking are the most common. Forward masking occurs when the masker is put before the test or pilot signal, and backward masking vice versa. The masking of a signal can occur as a result of a masker being placed above or below the signal in frequency (space). Upward masking refers to the masking of a signal by a sound of lower frequency, and downward masking by a sound of higher frequency than the signal. Egan and Hake (1950) found that upward masking was the more powerful of the

two when using a narrow band of noise as a masker and a pure-tone signal. This is generally held to be the case for all combinations of signal and masker. Before each type of masking can be described the critical band concept must be introduced.

1.2.3 Critical Bands and Critical Ratios

Fletcher (1940) proposed the idea that different frequencies produce maximal effects at different places along the basilar membrane. Each particular segment or filter responds to a limited frequency range, which he referred to as its critical band. He argued that the listener attends to the output of the filter with a centre frequency closest to the signal, and that background noise is only effective in masking the signal if it lies within the critical band of the filter being attended to. Fletcher's estimates of the critical band contained a systematic error due to his assumption that, within a critical band, an equal power of masking noise was required to mask a signal. This was subsequently found not to be the case. Later experiments suggested that a tonal signal may be approximately 4 db less intense than the auditory-filtered noise background before a listener fails to detect it. This corresponds to a post-filter signal-to-noise ratio of approximately 1:2.5.

Fletcher's estimates of the critical bandwidth are now referred to as critical ratios, and are on average 2.5 times smaller than current estimates of the critical bandwidth. It must be pointed out that when one refers to the bandwidth of "the" auditory filter, that it is not a constant value for all centre frequencies; bandwidth increases with frequency, more rapidly above 1 kHz.

It is not the purpose of this review to cover all the experiments and methods employed to determine the shape and bandwidth

of the auditory filter. The few methods reported here are included as they not only give estimates of critical bandwidth but also provide important information about the functioning of the auditory system.

1.2.4 Loudness Summation

One technique of deriving the bandwidth of the auditory filter is that of loudness summation. Scharf (1970) regards this as the most reliable method. Incidentally, loudness summation is one technique of bandwidth estimation that does not use masking. All experiments using loudness summation have shown that the loudness of a subcritical complex sound, of constant intensity, is independent of the bandwidth of the complex. Once the complex exceeds the critical band, then the loudness level increases. The theory behind this finding is that if the bandwidth of the complex is wider than a single critical band, then the loudness in each of the adjacent bands is added to give the total loudness perceived.

Graphs of loudness level plotted against bandwidth (for a given centre frequency) describe horizontal lines until the critical bandwidth is reached; then the loudness increases beyond the critical bandwidth. This can be illustrated for sensation levels of 20 dB upwards. There are notable exceptions to this, for example, when the complex is presented at a low sensation level (i.e. at around 10 to 20 dB SL) then the loudness is roughly independent of bandwidth. The loudness remains the same because if, for example, two critical bands are involved in processing the complex, the two component bands each contain half the energy of the original single band. When the complex is presented below 10 dB then as the critical band is exceeded, the level sinks below threshold as there is not enough

energy in any one or more component bands to sustain enough excitation to produce a sensation.

It must be pointed out that loudness summation experiments, in particular the earlier ones, were concerned more with loudness summation as a phenomenon than in estimating the critical band.

1.2.5 Auditory Filter Shape

The first attempt to measure auditory filter shape directly was carried out by Patterson (1974) using low-pass or high-pass noise maskers with a tonal signal. In order to overcome the (later thought to be false) assumption that the filter used for listening was the one centred on the tone, Patterson (1976) used notch or bandstop noise to determine the shape of the auditory filter. The signal was a pure tone, f_0, that was masked at frequencies above and below the tone by two broad noise bands with sharp cutoffs. He used the following equation to derive the auditory filter shape:

$$P_s = K\int_0^\infty N(f)W(f)df$$

where P_s represents the power of a signal at threshold; N(f) equals the power spectrum of the masker; W(f) is a weighting function applied to the power spectrum of a sound to determine the effective output of the filter at that frequency. K is a constant value.

An approximation of the filter shape is given by the equation

$$W(g) = (1 - r)(1 + pg)e^{-pg} + r$$

where g is the normalised separation from the centre of the filter, f_c, to an evaluation point f (where $g = |f - f_c| / f_c$. p determines the width of the passband of the filter. 1 + pg is a rounding factor, and r a dynamic range restrictor. Essentially, the shape of

the filter (as calculated from the above equation) is one of two opposing exponential functions separated by a rounded (rather than sharply peaked) top.

The theory states that as the noise band is moved towards the stimulus, threshold rises due to the increase of noise within the filter being used to detect the stimulus. The shape of the filter is then obtained by differentiating the tone threshold curve.

This method avoids a listener strategy known as "off-frequency listening". What the listener does in this case is to use the filter that gives the best signal-to-masker ratio, which may not be the filter centred at the signal frequency. Because the notch is symmetrical about the signal in notch noise masking the best signal-to-masker ratio should always be at the signal centre frequency. Obviously, to shift attention to either side of the centre frequency would just add more noise to the stimulus.

Patterson reported that the 3 dB bandwidth underestimates the equivalent rectangular bandwidth (ERB) of the derived auditory filters by approximately 14 %. The auditory filter is not rectangular but approximates to an asymmetrical Gaussian curve, with the steeper slope on the low frequency side. The ERB is a rectangular function with the same transfer function maximum and white noise energy transfer as the auditory filter, though obviously less wide.

Patterson concluded that the auditory filter shapes derived by this method revealed a sharply-tuned filter, and that the central portion of the filter shape can be approximated by a Gaussian curve. The auditory filter, however, is not a symmetrical function at most sound levels. Notch noise masking methods that place passbands symmetrically can reveal this asymmetry. If the stimulus frequency is kept constant whilst the notch is shifted up and down in

frequency, then any asymmetry would be revealed by a difference in tone threshold between equal upward and downward shifts. This is because differently shaped high and low frequency skirts would produce different signal-to-noise ratios for equal but opposite shifts in the notch band of frequencies.

Pick (1980) used comb-filtered noise (CFN) as a masker to study the level dependence of frequency resolution and consequently auditory filter shape. CFN, or ripple noise (as it is sometimes called), is produced by a pair of pseudorandom noise generators that create white noise which is added to a delayed version of the noise. This results in a sinusoidally varying spectral envelope of energy plotted against frequency (hence the name "comb-filtered"). By varying the delay the ripple (or peak density) can be altered. The peak density is the number of ripple peaks between zero and the frequency of the probe tone in frequency space. It was later developed to test the frequency resolution of subjects with cochlear hearing loss, and is reported in detail in Pick et al (1977). Pick (1980) illustrated some of the shortcomings of modelling the auditory system as a set of linear overlapping band-pass filters by showing that the auditory filter shape varies over level. We already know this from Egan and Hake (1950) and other similar experiments. But Pick produced more accurate data with a more sensitive technique.

One of the drawbacks of CFN masking is that unlike notch noise, its indications of filter shape become very inaccurate beyond 10 to 20 dB from the tip. This is because the weighting function, W(f), of the power spectrum is well defined around the tip, but the further one moves from the tip the less accurate this specification becomes (see Pick, 1980). A scale of level (in dB) plotted against log frequency gives a realistic indication of how bunched up the peaks of the noise are, and how much noise is interfering beyond 20 dB below

the tip. The power transfer function, W(f), was used (as in Patterson, 1976), to define the shape of the particular auditory filter being used to detect the tone. The results showed that the bandwidth increases at higher levels of noise and becomes more asymmetrical. These features are most obvious when the data are transferred into filter functions. A table of effective bandwidths (i.e. bandwidths of the equivalent rectangular filter) showed an increase with level for frequencies above 1 kHz; it also showed, interestingly, that at 8 kHz the effective bandwidth falls below the critical bandwidth (at low levels).

The results are important because many of the classical determinations of critical bandwidth and ratio did not really take level into account. Hence, if different experimenters use different levels of signal and masker, an unnecessary variability is introduced.

Moore (1983) provides a useful resume of past work in this review paper. His summary of filter characteristics include the following:

(1) The filter has a round top with skirts that are initially quite steep (around 100 dB per octave).

(2) The slope decreases beyond 25 to 30 dB of the tip.

(3) The ERB tends to increase with increasing centre frequency and also with increasing sound level.

(4) The low frequency skirt becomes shallower with increasing level, whereas the high frequency skirt becomes steeper.

Some of the above characteristics have direct implications for experiments reported in subsequent chapters. Where relevant these are referred to.

1.2.6 The Psychophysical Tuning Curve

An indirect but important method of determining frequency resolution is the psychophysical or psychoacoustical tuning curve (PTC). This method aims to measure the tuning (and hence selectivity) of a small group of hair cells on the basilar membrane. It is claimed to be the psychophysical correlate of the neural tuning curve which is determined by physiological methods. For example, Zwicker and Schorn (1978) used this method and compared data from classic masking experiments with PTC's and found good agreement.

In this method the signal level and frequency are fixed. The masker, which is either a sinusoid or narrow-band noise, is varied in frequency. The masker level required to just mask the signal is measured. The function produced, when plotted as frequency against loudness level (in dB) describes an inverted filter shape. Zwicker and Schorn concluded that PTC's are a useful tool for measuring the frequency resolution of normal and damaged ears.

1.2.7 Lateral Suppression

Important in frequency resolution studies, but not mentioned so far is psychophysical lateral suppression. It may be related to two-tone suppression when the response of a cochlear nerve fibre to a simple tone near CF can be reduced by a second tone at a higher or lower frequency. This appears to be a physical process present at the cochlear level of processing. Basically, each area of excitation on the basilar membrane resulting from a complex sound suppresses adjacent activity on the membrane. It may have implications for the frequency-resolving power of the auditory system in sharpening the tuning of the auditory filter.

Lateral suppression (which is not to be confused with inhibition, which is definitely a neural process) is not demonstrable in simultaneous masking, as the masker and signal in this case are being processed by the same channel. Houtgast (1972) argued, therefore, that any suppression effect would therefore be present in both tone and masker, leaving the signal-to-noise ratio unaffected.

Non-simultaneous techniques certainly seem to reveal effects similar to physiological suppression. PTC's using non-simultaneous masking show a greater frequency selectivity than those using simultaneous masking (Moore, 1978). It is not, however, possible to rule out the influence of true lateral inhibition occurring at the cochlear nucleus and higher levels of the auditory system in any psychoacoustical finding.

1.2.8 Conclusions

What then are the ramifications for speech perception and recognition from research findings of the frequency-resolving power of the auditory system? Briefly, the better the resolution of the system, the better the processing of speech sounds until temporal resolution suffers. Masking has a large effect on speech recognition. Each frequency component in a speech sound, if close enough, will have a masking effect on adjacent frequency components. Obviously, the better the resolution of the system in a temporal and frequency domain, the less masking there will be. But there is a theoretical limit, and improvement of one means a worsening of the other. The non-linearity of the high frequency side of the auditory filter also poses problems, which brings in the question of level as dealt with by Pick (1980).

1.3 Frequency Discrimination

1.3.1 Introduction

Frequency discrimination is the ability to detect changes of frequency as changes in pitch. The measure of this ability is referred to as the frequency difference limen (DL) (or sometimes the just noticeable difference (JND)). Pitch is a subjective attribute of a sound, and for a pure tone is monotonically related to the physical attribute of frequency.

There are two classes of theory that attempt to explain how we perceive pitch. These are the 'place' and 'temporal' theories of pitch perception. There are, in fact, differing versions of both of these theories. An understanding of the basis of these two theories is crucial to any discussion of experiments to determine frequency DLs.

1.3.2 Theories of Pitch Perception

1.3.2.1 Place Models

Different frequencies produce different patterns of activity on the basilar membrane. As frequency moves from high to low the pattern and location of maximum excitation progresses from the basal to the apical end of the membrane. Classical place theory, which was first argued convincingly by the German physicist Helmholtz, produces two assertions about this physiological activity. These are:-

(1) Some sort of spectral analysis takes place in the inner ear. This analysis results in different frequencies exciting

different places along the basilar membrane.

(2) The pitch of a stimulus is related to the position of the pattern of vibration that it produces on the basilar membrane.

The first of these assertions is universally accepted and is directly observable. The second is still partially a matter of contention within the field of hearing research.

1.3.2.2 Temporal Models

The temporal theory, or model, of pitch perception works by measuring the period of the stimulus waveform. This can be done by measuring the time intervals between nerve impulses. Hair cells tend to generate impulses when the displacement of the basilar membrane is at a specific phase. Not all the nerve fibres fire at a rate equivalent to the frequency of the stimulus. Nevertheless, they tend to be phase-locked to the stimulus, so that firings occur at the same phase of the waveform each time. More than one nerve cell fires in response to an incoming stimulus, and the phase-locking of several fibres over a time of up to 200 ms will give its frequency. This happens because the responses of adjacent nerve fibres are taken together and they add up to a response (or responses) on every cycle of a particular incoming frequency. Some fibres fire on one cycle some on another, but they all tend to fire at the same phase of the waveform.

The higher the frequency of the signal, and thus CF, of the fibre the less able it is to follow the waveform. Initially, this can be catered for by skipping alternate cycles and so on (the "volley" principle), but ultimately the refractoriness and imprecision of neural firing means that phase-locking breaks down

altogether at high frequencies. Palmer and Russell (1986) found that in guinea-pigs phase-locking declines beyond about 600 Hz until 3.5 kHz, above which it is no longer detectable. The frequency regions differ between species, though one might expect (from psychophysical data) the range to be higher and wider in man. Also, with a complex waveform phase-locking can occur at the period of the waveform even if the repetition rate of individual components is too high.

Palmer and Russell assert that the upper frequency limit of phase-locking will affect speech perception as some vowel second formants are as high as 3.5 kHz, thus precluding any higher formants from contributing to vowel quality. However, in a complementary study by Palmer et al (1986), as the above authors acknowledge, the use of the guinea-pig as a model for the temporal coding of speech is likely to underestimate the abilities of the human auditory nerve, i.e. man's frequency range is likely to be greater.

1.3.3 Experiments to Determine Frequency DLs

Many of the earliest experiments to determine frequency discrimination produced results which were unrealistically low at high frequencies, e.g. Shower and Biddulph (1931). Harris (1952) suggested that at very high and possibly very low frequencies, subjects in frequency DL experiments might well be detecting loudness rather than frequency differences. This is because at high frequencies, tones of equal amplitude sound fainter than tones at lower frequencies. Loudness differences can certainly be very marked for small frequency separations at high frequencies (and also under reverberant conditions).

Henning (1966), in a two-alternative forced-choice paradigm,

presented two tones which differed in both frequency and intensity. The task was to identify which of the tones was higher in frequency. The results show a marked difference from previous studies, Δf being larger than for fixed-amplitude stimuli by a factor of approximately 5 at 4 kHz. Frequency discrimination was shown to deteriorate rapidly above 3 kHz when loudness cues were not available. Henning concluded that frequency DLs at high frequencies involve loudness discrimination when fixed-amplitude signals are used.

Frequency discrimination used to be thought to obey Weber's law, which states that $\Delta f / f$ = a constant. Zwicker (1956) and (1970) attempted to explain the size of the frequency DL in terms of changes in the excitation pattern when the frequency is altered. An excitation pattern is a 'place' representation (of activity on the basilar membrane) where excitation (dB) is plotted against frequency (Hz). Zwicker maintained that a frequency change will be detected at the point of steepest slope (i.e. the low frequency side of the excitation pattern) when ΔE = 1 dB or more; ΔE being the change in excitation. The model predicts that the frequency DL at any given frequency should be a constant fraction (1/27) of the critical bandwidth at that frequency. (The steepness of the low frequency slope of an exciting tone is almost independent of both level and frequency). Experiments carried out in the present project test this assumption for frequency transition detection, using data from Chapter 4 to test the model.

A problem with the 'excitation pattern' model and Weber's law is that all the experiments to prove them used modulation methods. Henning (1966) had already shown the unreliability of this method. Moore and Glasberg (1986b) tested Zwicker's model by obtaining the slope of the auditory filter as well as intensity difference limens for frequencies between 0.25 and 8 kHz in subjects with unilateral or

bilateral hearing impairments. Their stimuli consisted of pulsed and modulated tones. The filter slope estimations were made by using a notched-noise masker method to measure the auditory filter shape (as outlined previously in Chapter 1.2.5).

Intensity difference limens were estimated for pure tones presented at 80 dB SPL; the aim was to determine the just detectable change in intensity in the low frequency side of the excitation. Frequency difference limens were also obtained by an analogous method to the one used for intensity.

To evaluate the excitation pattern model, Moore and Glasberg used obtained values of the slope and intensity DLs to predict the frequency DLs at various centre frequencies. Obtained and predicted frequency DLs were then compared. They found that for modulated tones the results were quite consistent with Zwicker's model (in both normal and abnormal ears). However, for pulsed tones, most of the obtained frequency DLs were smaller than predicted by the model.

Earlier experiments, such as Moore (1973a and b), attempted to bring together the place and temporal theories. Using short duration stimuli Moore (1973a) found a large increase in frequency DL for the lowest frequencies (125, 250 and 500 Hz); this also occurred to a lesser degree at longer durations. Between 4 and 5 kHz a sharp change was detected, the frequency DL becoming much larger (see Fig. 5.12). At 8 kHz, this deterioration tended to become less rapid. Moore predicted that both temporal and place mechanisms are in operation at the same time, and that the limits of performance are set by the more efficient mechanism at that particular frequency. It was already known that phase-locking of nerve impulses had not been observed above 5 kHz (Rose et al, 1967), and that at low frequencies the relative resolution of the basilar membrane is poor. The results were interpreted as showing a temporal mechanism as being the most

efficient mechanism in operation below 5 kHz, and a place mechanism above 5 kHz.

Moore (1973b) determined frequency DLs for narrow bands of noise as a function of bandwidth and centre frequency. The advantage of narrow-band noise is that it has 'built in' random-amplitude and random-phase variations. The most important findings were that the temporal model accounted for frequencies up to 4 - 5 kHz, and above this the place model appeared to be predominant (as in Moore's previous study). Both studies show a kink in the graphs between 4 and 5 kHz.

Goldstein and Srulovicz (1977) have shown, by mathematical modelling, that the optimum processing of auditory nerve interspike intervals can adequately predict the dependence of psychophysical frequency discrimination upon tone frequency and duration. They assume the ability of the auditory system to process the time intervals between successive nerve impulses, at the same time ignoring higher temporal dependencies.

1.3.4 Frequency Discrimination and Speech Perception

The auditory system's ability for frequency discrimination is greatly in excess of the requirements needed to detect changes in speech formants or intonation patterns. In fact the frequency DL in a normal ear is remarkably small and ranges from less than 0.3 % between 0.5 and 2 kHz to around 1.5 % at 10 kHz (Moore, 1973a; Evans, 1982) (e.g. 3 Hz at 1 kHz). Formants are concentrations of energy at resonance frequencies of the vocal tract. They describe broad spectral peaks when a vowel sound, for example, is represented in terms of amplitude and frequency.

The total range of fundamental frequencies that are normally

encountered in speech extends from around 60 Hz to about 500 Hz. The first four formants (which carry the majority of information for speech recognition) lie well within the band up to 4 kHz. This is important as the ear is most sensitive to changes in relative frequencies up to about 4 - 5 kHz, as previously discussed. Although formants give rise to peaks of energy in the spectrum that are relatively independent of the fundamental frequency, the first four formants nevertheless lie comfortably within a 4 kHz bandwidth.

It must be said however, that some speech synthesis systems generate the first six formants of speech sounds. So it may be misleading to completely discount the importance in speech recognition of formants above the first three or four. It may, conversely, also be useful to point out that the specified telephone bandwidth extends from 300 to 3400 Hz, and this tends to be quite adequate for the perception and recognition of speech. Also, Linggard (1985) states that between 3 and 5 formants are required for the generation of acceptable quality speech; the majority of synthesis systems use only 3 or 4. The higher formants in speech are concerned more with quality and, in synthesis systems, "naturalness". The lower formants, in particular F1 and F2 (and to a certain extent F3) in conjunction with articulatory cues, convey the identity of a phoneme or other such segment of speech.

Mermelstein (1978) measured DLs of formant frequencies for steady-state and consonant-bound vowels, and produced results very different from any discussed so far. The stimuli were generated by a synthesiser with the first, second and third formants (F1, F2 and F3) being variable, and the fourth and fifth formants (F4 and F5) fixed. The DLs for consonant-vowel-consonant (CVC) stimuli were found to be significantly larger than the DLs for steady-state vowels (i.e. vowels in isolation).

If one compares some of the values Mermelstein obtained to pure tone measurements (e.g. Moore, 1973a; Wier et al 1977) then one can see how large differences in thresholds are obtained with speech as opposed to non-speech stimuli. The mean DLs for F1 were 50 Hz for steady-state vowels and 49 Hz for a consonant context. The mean DLs for F2 were 142 Hz for steady-state vowels and 174 and 199 Hz respectively for /b/ and /g/ contexts. These values were obtained with centre frequencies of F1 and F2 in the same range as those in a similar study by Flanagan (1955a). (Flanagan used an F1 in the range 300 - 700 Hz, and an F2 in the range 1000 - 2000 Hz). Mermelstein's DLs are significantly higher than Flanagan's (i.e. between 9 and 14 % in percentage frequency terms in a CVC context). But even if one looks at Flanagan's F1 frequency DL of 15 Hz at 350 Hz (4.3 %), this is still much higher than any values discussed so far for non-speech sounds.

Haggard (1977) expressed no surprise that this is the case, and argued that we should expect a lower performance on formant discrimination than that for pure tones. An isolated vowel does not merely possess more spectral peaks than a pure tone; each formant is a resonance frequency of the vocal tract and is pulsed at the fundamental frequency. For the higher formants at least (usually F2 upwards), the impulse response of the auditory filters in these frequency regions are rapid enough to allow a decay in formant amplitude between each of the glottal pulses. Hence, in the higher formants, a certain amount of amplitude modulation occurs. Also, each formant peak has more than one harmonic component in it; one cannot view a formant as having a specific frequency but as containing a band of frequencies centred around the peak frequency. We know that the peripheral auditory system possesses enough selectivity to detect such spectral changes as those brought about by

a change in formant frequency. The problem is that the spectral changes can be so diffuse and so giving indistinct effects to the hair cells on the basilar membrane. Haggard says that "...the problem is not one of detecting that change in some frequency component has taken place, but rather of detecting a set of spectral amplitude changes from a limited possible set, analysing the consistencies in this set, and coding it as a formant change."

Frequency discrimination deserves a place in the current review for the compatability of DL with transition detection data for frequency. A comparison of the two tasks and results obtained from respective studies is included in the following section on FM. Also, an attempt is made in Chapter 5 to normalise data from a pure-tone frequency transition detection study (by the author) with well-known pure-tone frequency discrimination studies.

1.4 Frequency Modulation

1.4.1 Introduction

A frequency modulated (FM) signal is one whose instantaneous frequency changes in time. It is convenient to consider this as a constant (carrier) frequency (f_c) which is varied by a modulating waveform. As well as any complex waveform, FM can be analysed into a series of sinusoidal components. The depth of modulation is related to frequency swing, and if it is sinusoidal and small compared with the carrier, the FM wave can be considered as consisting of three components. These components are $f_c - g$, f_c and $f_c + g$ (Hz), where f_c is the carrier frequency, and g is the modulating frequency.

In discussing FM it is also important to consider amplitude modulation (AM) which is similar except that in AM the carrier's amplitude is varied in proportion with the modulating wave. In FM the amplitude remains constant whilst the frequency varies, in AM it is vice versa. Nevertheless these two types of signal can and do sound very similar in certain circumstances. As an AM and an FM wave both have identical frequency components then the only difference between them is in the relative phase of these components. Many studies in psychoacoustics have used as their basis the similarity between the spectra of FM and AM stimuli.

One example of their perceived difference lies in performed music. Vibrato, which is used extensively in vocal and instrumental performance, is periodic frequency modulation at a low rate and small extent. Although in practice it usually also contains amplitude changes. Vibrato tends to be perceived as such if the FM component is more pronounced. If the AM component is more pronounced then the listener will hear a tremolo sound.

Unlike many other areas in psychoacoustics, FM has received considerable interest in the speech literature. Many similar experiments have been done with FM stimuli in both psychoacoustical and speech research. Before going on to discuss various experiments it must be pointed out that non-repetitive stimuli such as frequency sweeps or transitions are also classed as FM signals, albeit only in one direction. Human speech contains unidirectional FM, for example in the formant transitions between consonant and vowel phonemes. The study of aural responses to transient stimuli is therefore necessary for understanding speech perception.

1.4.2 Frequency Transition Detection and Pitch Perception

The first issue to be considered is the threshold for detection of a frequency transition and what it depends on. Sergeant and Harris (1962) used glissandi with a linear rate of frequency change to find human sensitivity to unidirectional frequency modulation. Glissandi were always started at the same frequency (1500 Hz), and subjects were asked to judge whether the pitch of the stimulus rose or fell. They found that the total frequency change which must occur for a sensation of glissando to appear was approximately 5 Hz at 1500 Hz, which is the same as reported frequency DLs at that frequency. They plotted the threshold of glissando rate as a function of stimulus duration and found that the rate decreases as stimulus is increased up to durations of 10 s. Between 10 and 25 s, glissando sensitivity starts to break down which was shown by a threshold increase. Collins and Cullen (1978) investigated tone-glide detection as a function of both intensity and duration. They found that thresholds were affected by duration and extent of frequency change (as well as signal-to-noise ratio) and in a parallel manner to

steady-state signals. Tones were presented at various intensities against a constant noise background, and were in the F1 and F2 frequency regions of speech sounds.

Nabelek and Hirsh (1969) also performed a speech orientated psychoacoustic experiment on frequency transition discrimination. They used durations, frequency changes and frequency regions particularly related to those values encountered in speech perception. They found that discriminability for a small frequency interval (9/8), was poorest when the transition duration was shortest (10 ms). For nearly all other frequency intervals the best discriminability was obtained for transition durations of around 30 ms. This value is in the range of frequency transitions in speech, most especially formant frequency regions. The best discriminability of frequency changes for transition durations of around 30 ms is a general property of hearing, and is not just a feature of speech perception.

Tsumura et al (1973) performed experiments to obtain thresholds for detection of frequency transitions whilst varying the duration of the transition, delay of transition onset, and duration of the final part of the stimulus (which had a fixed frequency). As in above experiments frequency was changed linearly. Subjects were asked to decide which burst in a pair contained a frequency transition. They found that as the delay of transition onset increases, the threshold of frequency transition shows a tendency to decline. Also, the pitch difference between the initial and final steady frequency segments is a main cue for the detection of a frequency transition if the duration of the transition is less than 5 ms. Following on from their results the authors suggested that as long as the terminal steady frequency segments have sufficient durations before and after the transition, that pitch difference is a dominant cue for the

detection of frequency transitions. This was said to be so even if the duration of transition is relatively long.

This brings us to the question of what pitch an observer perceives an FM stimulus to have. Brady et al (1961) performed an experiment in which subjects were asked to match a stimulus of rapidly varying resonant frequency to a comparison stimulus of steady frequency (the comparison being manipulated by the subject). The stimuli were characterised by different initial and final frequencies (in the 1000 - 2000 Hz range), different durations of frequency transitions (in the 20 - 50 ms range), and different numbers of 100 cps excitation pulses (from 2 to 6). The transitions either took up the entire stimulus or there were covarying durations of initial and final steady-state. The repetition rate of pulses was always 100 cps, which is in the range of glottal frequency for male speech. Also, the frequency transitions were typical of those found in the second formant in speech. Subjects showed a consistent tendency to set the resonant frequency close to the terminal value of the stimulus. This tendency became stronger as the rate-of-change of frequency was increased and also when an excitation pulse occurred after the stimulus had reached the terminal frequency.

McClelland and Brandt (1969) went further than Brady et al by attempting to determine the pitch of frequency-modulated sinusoids rather than just unidirectional modulation. The depth of modulation was a constant 100 Hz whilst the centre frequency of the sinusoid was varied between 300 and 3000 Hz, and the modulating frequency between 20 and 2000 Hz. Subjects were instructed to match the stimulus nonsimultaneously with a pure tone. They found that pitch matches to components in the stimulus depended upon their frequency spacing, i.e., when individual components exceeded critical bandwidth they were perceived separately. When the spectral components did not

exceed the critical bandwidth, pitch matches were made near the centre or the ends of the band of spectral energy. When the level was increased from 20 to 50 dB SL, with greater than critical bandwidth component spacing, an additional stimulus component became audible, and a pitch match was made. This would presumably have been present but below threshold at 20 dB SL. The authors also noted the markedly differing sensations produced at different modulation frequencies. These have been shown to depend on modulation frequency, frequency deviation, centre frequency and sound pressure level (Fastl, 1983).

1.4.3 Differences in Sensitivity to Upward and Downward Sweeps

An important finding, and one that recurs in many FM studies, is that subjects respond differently to rising and falling transitions. Brady et al found a stronger tendency for subjects to match the terminal frequency of the stimulus when the transition had been rising rather than falling in frequency. Gardner and Wilson (1979) found that subjects were more sensitive to upward than downward frequency sweeps. They contrasted their results with Tsumura et al (1973) and Nabelek and Hirsh (1969). Tsumura et al found a small difference but in the opposite direction, and Nabelek and Hirsh apparently none at all. This, Gardner and Wilson felt, was partly due to the fact that the task was more one of detecting a frequency difference between stimuli rather than the presence of a transition. This is a valid criticism of many of the above experiments, as they do seem, at times, to confuse transition detection with frequency discrimination.

Further support for greater sensitivity to upward sweeps has come from Cullen and Collins (1982) in some experiments to measure

the effects of rate of frequency change. A subsequent study designed to investigate the effects of background noise on tone-glide perception yielded further supportive results (Collins and Cullen, 1984). Horst (1982), using simple approximations to second formants found subjects to be slightly more sensitive to upward frequency sweeps than downward sweeps, but the difference was not significant.

Finally, Lacerda and Moreira (1982) also produced results indicating greater sensitivity to upsweeps. They explained the results of a masking experiment in terms of the interaction between the residual excitation that occurs shortly after the offset of a stimulus, with lateral suppression. The latter takes place during simultaneous masking when masker and signal are being processed in the same channel (see Chapter 1.2.7). Lacerda and Moreira argued that if a masker followed a frequency sweep quickly enough in time then lateral suppression could occur. The resulting 'sharpening' of the signal would, they argued, have a greater effect on upward sweeps and make them more detectable than downward sweeps (due to the shape of the suppression curves).

1.4.4 Adaptation Studies with FM stimuli

In the 1960's a good deal of physiological work was done on cells in the auditory cortex that respond specifically to FM stimuli. This led on to psychoacoustical studies on feature detection. Kay and Matthews (1972) produced selective adaptation with a sinusoidally frequency-modulated tone, i.e., they elevated FM detection thresholds. Gardner and Wilson (1979) produced further evidence for FM feature detectors using a variety of adapting stimuli. Using both sinusoidal FM and single linear frequency sweep stimuli they found that only upward sweep stimuli were able to adapt upward sweep

signals, and only downward sweep stimuli adapted downward sweep signals. In other words they provided evidence for separate channels specific to upward and downward FM. The adapting sweeps used were similar in frequency regions, range of frequency excursion, and rate-of-change of frequency to some second formant transitions in speech.

Diehl (1975) looked at the effect of selective adaptation on the identification of certain consonant-vowel (CV) syllables. He found that there was a net change in the number of stimuli identified as /b/ or /d/ due to different adapting stimuli. The shift in identification, when it took place, was most pronounced when the stimuli lay near known phonetic boundaries.

Tartter and Eimas (1975) also looked at the issue of feature detection in speech perception. Like Diehl they used CV stimuli, which varied only in the starting frequencies and directions of second and third formant transitions. They found that the most effective adaptor of speech sounds were similar speech sounds rather than any of the several isolated components of the entire pattern.

The most important implication of the above two studies is that the adaptation of speech sounds may be due to feature analysers similar to those mentioned for FM detection. Also it may be that FM detectors are a subset of a whole group of feature detectors that serve both the perception of speech and other complex sounds.

1.4.5 The Role of Formant Transitions in Speech Perception

When a consonant phoneme is placed before or after a vowel, transitions in frequency occur in the formants, most notably in F1 and F2; this takes the formants from their (steady-state) vowel position to a different region of frequency space. As previously

stated, frequency transitions are frequency modulations in one direction and therefore must be considered as FM studies.

Delattre et al (1955) looked at the effect of varying the locus (or frequency position) of the first and second formants of the voiced stop consonants /b/, /d/ and /g/. They also varied the duration of a succeeding vowel's steady-state. They found the most important factor in discrimination was the locus of the F2 at the beginning of the utterance. With a fixed F1 and F2 locus it was found that subjects heard successively /b/, /d/, /g/ and finally /d/ again as the steady-state of the F2 was varied from 2520 down to 1200 Hz. This clearly demonstrates the importance of the frequency of the second formant transition in distinguishing one stop consonant from another.

O'Connor et al (1957) attempted to determine the acoustic cues necessary for the recognition of the semivowels /w/ and /j/ and the glides /r/ and /l/ in their initial position before vowels. The distinctions amongst the four phonemes proved to be dependent primarily on their F2 and F3 transitions (in both direction and extent). Also, the duration of steady-state onsets, i.e. the period of no frequency change, was also shown to be important.

Liberman et al (1954) had previously shown that direction and extent of F2 transitions was important for distinguishing between stop consonants (with a following vowel). Ainsworth (1968), however, found transition duration (in the range 100 to 200 ms) to be relatively unimportant for the perceptual discrimination of semivowels, as did Liberman et al (1956). Greater importance was attached, in Ainsworth's case to the locus of F1 and it's 'shape' of transition.

Lehiste and Peterson (1961) also showed the importance of formant transitions in the discrimination of glides and diphthongs.

Lindblom and Studdert-Kennedy (1967) looked at the role of formant transitions in the recognition of vowels. They used the vowels /I/ and /U/ in a /j - j/, /w - w/ or isolated vowel context. They found that vowel stimuli which were identical in all other respects than their formant transitions were not perceptually equivalent. They tentatively suggested from their results that it was not just the formant pattern in the vowel's steady-state that determined it's perceived identity but also the direction and rate of adjacent formant transitions.

Stevens and Klatt (1971) looked at the role of formant transitions in the voiced-voiceless distinction for stops. Their stimuli were the stops /da/ and /ta/ produced by a speech synthesiser. They found that the voice onset time (VOT) and the degree of first formant transition are important cues for the voiced-voiceless distinction for stop consonants. (VOT is the time from release of closure to the onset of vocal-cord vibration). Wang (1959) in a study designed to investigate the perceptual cues used for stops, (this time) in final position, agrees with this view. Wang suggested that "...the value of formant transitions as perceptual cues for consonants depends on the magnitude of the formant movements."

Holmgren (1979) attempted to determine whether formant frequency target or rate-of-change of formant frequency is more important in the perception of the identity of a vowel. Using stimuli designed to be perceived differently if the former or latter is more important, she obtained results that were by no means clear cut. There was, however, a tendency towards formant frequency target (rather than rate-of-change of formant frequency) as being the more important source for vowel identification.

There is some support for this view from Chistovich et al (1982)

who argued that listeners identify formant frequency dynamics by trajectory endpoints rather than rate-of-change. Also, Gay (1970) asserted that American English diphthongs (namely /Oi/, /ai/ and /au/) "...are characterised primarily by an invariant rate of formant frequency change".

The position is still not entirely clear, though the current weight of opinion, although not discounting the contribution of rate-of-change of frequency as an identification cue, is in favour of formant frequency target as being the major determinant.

1.4.6 Frequency Discrimination and Frequency Transition Detection: Differences and Similarities

Frequency discrimination, as stated earlier (see Chapter 1.3), is the ability to detect changes of frequency as changes in pitch. These changes in frequency actually refer to non-simultaneous sounds. If these sounds were presented simultaneously then the task would be one of frequency resolution (see Chapter 1.2). Frequency transition detection is the ability to detect a change in the frequency of a sound, or one of its components, that is sweeping in frequency.

The most important aspect of this area of discussion of the two tasks is how similar they are in terms of threshold values rather than how they have been conceptually confused. It is certainly difficult to make direct comparisons between data for both tasks due to the attributes of frequency sweeps. These include the amount of frequency change, initial and final frequency values, rate-of-change of frequency and duration of frequency transition. Also, the sweep can be linear or non-linear over time over the chosen frequency scale.

Horst (1982) carried out experiments on both the detection and

discrimination of frequency sweeps. Frequency sweep discrimination is the ability to judge two signals as being perceptually distinct due to differences in their rate of frequency change. Using band-pass filtered pulse trains Horst produced bell-shaped (parabolic) sweeps with the same initial and final frequencies. With these stimuli he obtained frequency sweep detection thresholds of between 0.4 and 2 %, and sweep discrimination thresholds of between 1 and 3.5 %.

Both Horst, and Lacerda and Moreira (1982), suggested that the auditory system treats frequency transitions (formant or otherwise) as a sequence of short stationary segments. This would imply that the mechanism responsible for frequency discrimination and transition detection is the same in both cases, and consequently that the two tasks should produce identical results. Of course, the latter assertion depends on the frequency excursion, rate of frequency change and the duration of transition.

Workers who do express an opinion believe frequency transition detection and frequency discrimination to be related abilities, and that the latter is the more sensitive of the two. How to compare the data is the most complicated issue as, for the above detailed reasons, it is difficult to decide what the equivalent stimulus conditions are. It was easier for Horst to do this with time-varying stimuli being employed for both of the tasks he was comparing, but not for the present attempted comparison where only one of the tasks involves a dynamic stimulus.

The studies where a real comparison may be possible are those that use pure-tone stimuli. As Hoekstra (1979) points out in reviewing frequency discrimination studies "...for complex tones there is not just a single jnd for frequency. The result of a frequency discrimination measurement depends on the composition of

the complex tone and on the variation applied." This argument can be equally applied to frequency transition detection studies. In Chapter 5 a comparison is made between pure-tone frequency transition detection data (obtained by the author) and two well-known pure-tone frequency discrimination studies in an attempt to make a realistic assessment of the differences.

1.4.7 Conclusions

What the above speech experiments (in Chapter 1.4.5) show is that the auditory system makes great use of frequency transitions in the perception and discrimination of speech. What is yet to be determined is clear physiological evidence whether there are feature detectors specific to FM (and all other types of stimuli) in the auditory system.

The psychoacoustical and speech literature relating to frequency modulation clearly shows that different responses are achieved depending upon whether a subject perceives a stimulus to be 'speech-like' or not. Liberman et al (1967) suggested that speech-like sounds are processed in a different way. The responses of subjects in the experiments of Brady et al (1961) are enlightening inasmuch as the authors tried to present 'speech-like' stimuli, yet no subject perceived them as such. The issue of whether there is a speech mode of perception is dealt with below in Chapter 1.6.

1.5 Intensity Discrimination

1.5.1 Introduction

Intensity discrimination is a measure of the ability to detect changes in intensity, and is usually defined as that intensity difference (expressed in dB) which produces 75 % correct responses in a two-alternative forced-choice (2AFC) procedure. In this review section quantitative measures of intensity discrimination (DLs) and how intensity is coded in the auditory system will be discussed. The former will be dealt with first.

1.5.2 Measurements of Intensity DLs

One of the earliest studies of intensity DLs was carried out by Riesz (1928) to determine the discrimination of pure tones. He mixed the outputs of two oscillators which were set at slightly different frequencies so that periodic smooth intensity changes were present in the signal. The data for this amplitude modulation technique show a DL of 1.5 dB at 20 dB SL, which drops to 0.3 dB at 80 dB SL, at a frequency of 1 kHz.

Miller (1947) determined intensity DLs for white noise. He used continuous noise with a relatively uniform spectrum to which an increment was periodically added. Miller plotted the increments which could be heard 50 % of the time against the sound level of the noise.

The data indicate that for an intensity of 30 dB SL or more, the relative differential threshold (ΔI / I) is approximately constant for wideband noise, i.e., Weber's law holds. (Weber's law states that $\Delta I = k.I$ where I is the size of the JND, k is a constant, and I

is intensity). At the highest intensities measured the intensity DL was approximately 0.41 dB. Miller concluded that "...all these studies indicate that the difference limen for intensity is of the same order of magnitude for noise as it is for tones, at least at the higher levels of intensity. At the lower intensities the discrimination for a noise stimulus maybe somewhat more acute than for tones."

Subsequent experiments to those of Miller and Riesz have gone further in attempting to explain how subjects perform intensity discrimination tasks. Viemeister (1972) determined intensity DLs for pulsed sinusoids and suggested that, with this method, aural harmonics were used as additional cues for detecting a change in intensity.

Three types of noise spectra were presented with a sinusoidal pulse. These were a low-pass noise (with a cut-off 3 dB down at 800 Hz), a bandpass noise centred at the sinusoid's second harmonic, and a high pass noise (with a cut-off at 1900 Hz). The frequency of the sinusoidal pulse was 950 Hz. The effects of these noise spectra on the slope of the masking function for the intensity discrimination of pulsed sinusoids were examined.

Viemeister found that when frequency regions at and above the second harmonic of the signal were masked, Weber's law held. His interpretation of this was that subjects used intensity information at aural harmonics for the discrimination of tones, and that was the reason for the deviation from Weber's law shown by the intensity discrimination of pure tones. Viemeister proposed a simple model which states that Weber's law applies for the fundamental and each aural harmonic, but the Weber fraction ($\Delta I / I$) decreases as successively higher harmonics become audible.

Moore and Raab (1974) repeated Viemeister's experiment, but

using tonal maskers at aural harmonics, in an attempt to eliminate the 'near-miss' to Weber's law as Viemeister had done. They criticised Viemeister for using noise levels more intense than seemingly necessary, to mask the aural harmonics. They used lower levels of masking stimuli in an attempt to verify Viemeister's experiment.

From their results they concluded that the explanation of the 'near-miss' to Weber's law in terms of aural harmonics is not appropriate. Moore and Raab found it very difficult to impair intensity discrimination at high levels, so they used noise maskers and produced results similar to Viemeister (1972). With tonal maskers they could only eliminate the 'near-miss' as Viemeister was able to do when the level of the second harmonic well exceeded what would normally be present in a pure tone.

In concluding their study Moore and Raab questioned why Weber's law should hold for the intensity discrimination of pure tones in any case. Weber's law holds for the intensity discrimination of noise due to the statistical properties of noise itself. This is a very pertinent point as Weber's law is an empirical law which probably relates to underlying compression mechanisms; there seems to be no biological reason for it to be an exact 'law'.

Viemeister (1974) determined intensity DLs for noise in the presence of band-stop noise. The effects of stop-bandwidth and relative level of band-reject noise on the intensity discrimination of broad-band noise were examined. Discrimination thresholds for 200 ms bursts using a two-interval-forced-choice (2IFC) tracking procedure were obtained. For each trial two noise bursts N_0 and $N_0 + S_0$ were presented 500 ms apart, the level of increment (S_0) being varied from trial to trial. $\Delta I / I$ was plotted against N_0. The results showed that Weber's law holds over at least an 80 dB

range.

Viemeister found that the addition of a band-reject noise generally caused an increase in the Weber fraction, and attributed it to a reduction in the effective bandwidth used by subjects in discriminating the signals.

Moore and Raab (1975) also determined intensity discrimination for noise bursts in the presence of a bandstop (or band-reject) background. They also wanted to investigate the effect of varying the effective bandwidth available to the subject to see if differences in performance for stimuli of short and long duration could be detected. They found that for long duration stimuli (250 ms) there was a reasonable approximation to Weber's law in the absence of background noise. The performance at short durations (10 ms) was worse than at long durations without background noise. When background noise was added, performance worsened in both cases. Thresholds also increased at higher levels, but only by a small amount.

1.5.3 Intensity Coding

None of the above results can be adequately explained without an understanding of how intensity is coded in the auditory system. In order to further this understanding at least two questions need to be answered. The first, which arises directly from the experiments just described, is why Weber's law holds for noise bursts but not pure tones. The second question is how is the auditory system able to detect changes in intensity over a range in excess of 100 dB when the average nerve fibre reaches saturation in firing within 40 dB of its threshold?

It was thought for many years that the wide dynamic range of the

auditory system was due to the fact that there were different groups of fibres having widely differing thresholds. Kiang (1968) attributed this misconception to the pooling of data across cats that had considerable differences in their individual results. Kiang suggested that auditory nerve fibres could be thought of as a pool of elements where the activity of single elements can be related to characteristic frequency (CF) or other factors that determine the absolute rates of discharge.

Unfortunately, this still did not answer the question as to how a sound above (approximately) 60 dB SPL could be coded for intensity. In fact it made things more more difficult to account for and understand.

All auditory nerve fibres fire spontaneously with fairly widely differing rates. They show frequency selectivity and phase-locking. Their CF is the frequency at which their threshold is lowest. A single nerve fibre is assumed to derive its input from a particular part of or 'place' on the basilar membrane. These fibres collectively signal changes in intensity by an increase or decrease in firing rate. Each auditory nerve fibre can make hundreds of connections with cells in the cochlear nucleus, which is one stage higher in the auditory processing chain which runs from the basilar membrane level to the auditory cortex. What happens is that fibres in the auditory nerve converge and cross-connect in the cochlear nucleus. This enables one fibre to communicate and combine with many others at a higher level. Fibres with the same CF combine with each other via phase-locking and multiple cochlear nucleus connections, to signal intensity changes.

Sachs and Abbas (1974) observed discharge rates of single auditory-nerve fibres in cats in response to tone bursts. The tone bursts were varied in intensity to investigate the effects of level

upon discharge rate. They found that for tones at a fibre's CF, that the rate increases rapidly over a range of 20 to 30 dB above their thresholds. For higher stimulus levels, however, a range of behaviours was observed. Some fibres saturated completely at 30 dB above threshold whereas other fibres produced a noticeable kink in the rate-level function at 20 to 30 dB above threshold. The function started to flatten out at this level and hence the fibres were becoming less sensitive. The rate in this latter group increased more gradually over a further 30 to 40 dB. This 'sloping saturation' suggested that the auditory system's wide dynamic range may be accounted for by a population of cells that become less sensitive at higher levels, yet are still capable of signalling intensity changes.

Palmer and Evans (1979) investigated these fibres with a larger dynamic range, but found that only 9 % of the fibres investigated in the auditory nerve of the cat had dynamic ranges in excess of 60 dB, and only 3 % greater than 70 dB. They concluded that although there were fibres that were unsaturated at high levels, their usefulness was extremely doubtful. This is because of the proportion of the fibres being small, and the fact that they all showed a substantial reduction in slope by 60 dB SPL meant that there was substantial attenuation above this level. Thus, even though high levels could be coded in this way, the DL would inevitably become much greater. This once more dislodged the idea that intensity can be coded at high levels by the firing rates of auditory nerve fibres. This contrasts quite markedly with the views of Carlyon and Moore (1984) (reported in more detail below), who suggest that a decrease in their relative intensity DL thresholds between approximately 60 - 85 dB SPL could be due to the activity of this relatively small population of high threshold fibres.

Evans (1981), however, suggested an alternative possibility to

the use of wide dynamic range fibres. Evans stated that "...because the fine time structure of the discharge patterns of cochlear fibres reflects their filtering (for frequencies up to 4 - 5 kHz) for all sound levels regardless of saturation of their mean discharge rate, this could be used to encode the 'place' information. If the higher levels of the auditory system could in fact extract this information, this would be a very robust 'place' representation of complex spectra as those of speech."

There is also an additional mechanism to take into account. Fibres of neighbouring CFs can 'help out' in coding at high levels. Although at 60 dB SL virtually all the fibres at a particular CF are saturated, those in neighbouring areas are not. A change in intensity could then be signalled by a spread in the excitation of the stimulus on the basilar membrane. This idea is not new, and has received a certain amount of coverage in Chapter 1.2 on frequency discrimination.

Moore and Raab (1975) found that the width of the stopband (of a continuous bandstop background noise) had an effect on intensity discrimination of broad-band noise bursts. They found the greatest impairment to discrimination was with the narrowest bandstop. Performance improved with increasing stop bandwidth and was best when no background was present. This confirmed that information at frequencies other than the CF is important for discrimination. It is only a short step from there to saying that the information from firing of fibres with neighbouring CFs are used, especially at high levels, to aid intensity discrimination.

This certainly seems to be possible for pure tones which have stable edges of excitation. Noise bands, however, are a little problematic to this approach, as their edges are very unstable due to fluctuations in amplitude. Yet Moore and Raab (1975) found that even

though they disrupted information on both sides of a noise band, in the presence of continuous bandstop background noise, intensity discrimination was still quite good, even at higher levels.

Viemeister (1974) came to the same conclusion as Moore and Raab inasmuch as he assumed that the subject makes a broad-band observation in the absence of band-reject noise. When the noise is added the effective bandwidth is decreased, and so the subject is forced to listen to noise components falling primarily within the stopband of the masker. Hence the performance worsens.

Some interesting results were produced by Evans and Palmer (1975) who recorded the responses of cochlear nerve and nucleus fibres to signals in the presence of bandstop noise. They found that the dynamic range for cochlear fibres was unchanged by the addition of the bandstop noise masker, and that even when (as expected) the neurones in the auditory nerve showed saturation at high levels, 59 % of neurones recorded in the dorsal cochlear nucleus responded to signal level changes over a wide intensity range.

Carlyon and Moore (1984) measured intensity DLs of tones between 500 and 8000 Hz in the presence or absence of bandstop noise. Stimulus levels used were between 20 and 85 dB SPL. The onsets and offsets of the stimuli were masked to see if they provided cues to aid performance of the task. They found that intensity discrimination worsened at high frequencies and (their) mid-range levels, in particular for short duration tones in bandstop noise. They concluded from this that spread of excitation seems to be an important cue for the task (which is implied in the results of Moore and Raab, 1975). This is because the bandstop noise can mask regions adjacent to that of the stimulus frequency which might otherwise detect such a spread.

Taken together, the above results suggest that no single mechanism has yet been identified that can account for the coding of intensity in the auditory system. Evidence suggests a combination of excitation pattern spreading, phase-locking cues, changes in firing rate, high threshold neurones with sloping saturations and perhaps other mechanisms which all play a part in intensity discrimination.

Evans and Palmer (1980) reported evidence to suggest two overlapping subpopulations of cochlear nerve fibres, those with discharge rates below 15 spikes per second and those above. There was a clear trend in which fibres with the highest spontaneous rates have the smallest dynamic range. Members of these two subpopulations were found to have overlapping distributions. Fibres with dynamic ranges in excess of 60 dB were only found in substantial numbers in the population that had low spontaneous discharge rates. The population below 15 spikes per second formed 35 % of the total. They may contribute in some way to intensity coding albeit perhaps in a minor way. Rouiller et al (1983) found that 38 % of extracellularly recorded fibres in the medial geniculate body of the cat had a mean dynamic range of 60 dB. The other 62 % of the fibres had dynamic ranges in excess of this level.

Winslow et al (1987) reviewed recent work with regard to rate-place representations of pure tones in quiet and noise backgrounds. Their main conclusion was that the small population of low spontaneous rate high-level saturation fibres are crucial for the maintenance of theories of intensity coding that are based on rate-place representations (over a wide range of levels). As this population is so low (Palmer and Evans, 1979; Evans and Palmer, 1980; Rouiller et al 1983) then some sort of weighting mechanism is necessary to boost their relative contribution to the signalling of intensity changes.

Winslow et al proposed a simple level dependent weighting function based on the principle of direct path inhibition. Briefly, their model works if one assumes that the low spontaneous rate fibres are less distally located (synaptically within the dendritic tree) than similar CF fibres with high spontaneous rates. This enables, at high levels, the activity of low rate fibres to inhibit those of high rates, thus weighting the formers' response higher (in terms of population size) than the greater numbers of high spontaneous rate (smaller dynamic range) fibres.

Perhaps the greatest puzzle of the intensity coding problem however, is the comparative constancy of DL at different SPLs in spite of the very different modes of action of the various mechanisms discussed above.

1.5.4 Intensity Discrimination and Speech Perception

Compared to other areas of auditory psychophysics, such as frequency discrimination and modulation, intensity discrimination has received much less attention in speech research. There are, however, a few relevant experiments that can be discussed.

Flanagan (1955b) undertook an exploratory experiment to determine just discriminable differences in the intensity of a synthetic vowel sound. The synthetic vowel had four formants of fixed frequency. The listening test consisted of an intensity judgement between two stimuli; the comparison stimulus was identical to the standard apart from the fact that it differed in intensity by a known amount. The second stimulus of the pair was incremented in 0.4 dB steps (though in some cases it was not incremented at all). Each stimulus component had a duration of one second and data were collected for three different time spacings (0.5, 1.0 and 1.5 second

spacings).

The total percentage of 'different' votes (i.e. judgments that the second stimulus of the pair was not identical) was plotted against the attenuation of the second component stimulus for all three time spacings. The 50 % 'different' point was taken as the value of the difference limen. Over all the tests the median intensity DL was 1.5 dB, and never less than 1.2 dB.

As this was only a pilot study the data should be viewed with caution, and as only one vowel sound was used, its limitations should be realised. In comparison to pure tone and wideband noise data, discrimination seems to be a little poorer for the synthetic vowel which was presented at 70 dB SPL (conversational speech level).

Flanagan (1957), using approximately the same method as above, tested intensity discrimination of the second formant of the vowel /{/. ({ is the symbol chosen from Wells, 1987 to represent the front short vowel from, for example, the word 'act'). A standard sound followed by the same sound with the second formant increased or decreased formed the stimulus pair. Subjects were asked to judge if the 'quality' of the second component of the stimulus was the same as or different to the first. There were two different tests, the first using a constant fundamental (120 Hz) and the second a linearly inflected fundamental (95 to 105 Hz).

As in Flanagan (1955b) the percentage of 'different' judgments were plotted as a function of change in second formant amplitude. For the fixed fundamental condition the curves were symmetrical, and so the percentage of 'different' judgments were not much influenced by the polarity of the change in amplitude of the second formant, but were mainly dependent on the magnitude of the change. In the variable pitch condition the results were very similar. Both curves yielded a difference limen of approximately 3 dB, if the DL is

regarded as the change in second formant amplitude noticed 50 % of the time.

Flanagan's justification for using the vowel /(/ is that the formants of this vowel are nearly equi-spaced in frequency and that the second formant lies near the centre of the frequency range usually containing second formants. Considering that the vowel /(/ represents one of the more favourable conditions for detecting small changes in the amplitude of the second formant, Flanagan reasoned that with different vowels one would expect DLs to be not much smaller and probably larger.

Flanagan did this experiment to provide an estimate of the maximum precision necessary in transmitting data on the amplitude of the second formant in vowels. He concluded that "...it should not be necessary to transmit data specifying the amplitude of the second formant within an accuracy exceeding ± 40 %." (He arrived at this value by putting the value 3 dB into the equation $20\log_{10}(A \pm \Delta A)/A)$.

There have been other investigations relating to intensity changes in speech (as well as those of Flanagan) with regard to formant amplitude, and overall intensity, discrimination. Fry (1955), for example, investigated word stress (i.e. emphasis) and found that the syllables in a list of words that were judged to be stressed had the longest durations and the greatest intensities. When Fry compared the two factors he found that duration was the more important of the two. In a subsequent study Fry (1958) found that syllables with higher fundamental frequencies were perceived to be stressed. So, intensity differences between two syllables can be completely absent but one can still be judged to be stressed.

As far as vowels are concerned, intensity changes tend to occur during coarticulation, i.e. when formant frequency transitions are taking place. Intensity tends to be constant during the steady-state

portion of a vowel. Before the detectability of simultaneous frequency and intensity changes can be investigated, however, sensitivity to intensity transitions alone must first be considered. Chapter 7 is an attempt at such a study.

1.6 Evidence For and Against a Speech Mode of Perception

In order to do justice to such an important and contentious topic such as this would need a sizeable book. This is obviously not possible in the present case, instead some of the arguments and demonstrations for and against a speech mode of perception are briefly discussed. The intention is to present some of the evidence rather than argue the case for one side, though most of the protagonists in this field turn out to be one-sided.

The only convincing way to argue against a speech mode of perception is to be able to show that in all imaginable cases a correlated response between psychoacoustic and speech stimuli occurs, or at least for all conditions of speech production. If one provides one, or perhaps several, cases to the contrary then that is arguably sufficient proof that such a mode exists. It is not necessary that such a speech mode must continuously be in operation or be predominant for it to be proved to exist. It may be that it only operates or predominates when speech cues are at their most powerful (see Chapter 6).

One of the prime (and often used) indicators of the existence of a speech mode (which for the time being we will assume does exist) is the presence of categorical perception in a speech discrimination task. Repp and Williams (1986) define categorical perception (or tendencies) as "...nonlinear stimulus-response mappings of formant frequencies, non uniform response variability across each continuum, and peaks in formant frequency distributions."

Mattingly et al (1971) performed a set of experiments (which are reported in detail in Chapter 6) to test listeners discriminative abilities with regard to formant transition rate. They found that for simple speech stimuli discrimination functions peaked at phonetic

boundaries (these being implied by previously constructed identification functions). Where "chirps" were used (i.e. isolated formant sweeps) discrimination functions were random. This, they concluded, provided clear evidence for a speech mode of perception which was not, according to them, switched on all the time during speech events.

Liberman (1984), not surprisingly, takes the same view. He argued that listeners perceive speech automatically as a phonetic structure, and that they do not translate incoming acoustic information from an auditory base; he also stated that "...taking stop consonants and their dynamic formant-transition cues as a particular example, I note that listeners are not aware of the transitions as pitch glides (or chirps) and also as a (support for) a stop consonant. Listeners are only aware of the stop. Yet these same formant transitions <u>are</u> perceived as pitch glides (or chirps) when on the nonspeech side of a duplex percept, for example, they do not figure in the perception of a phonetic segment."

Pisoni (1983) views things in the same way, and sums up the situation by saying "Differences in perception between speech and nonspeech signals were assumed to reflect two different modes of processing; a "speech mode" and an auditory mode. Despite some attempts to dismiss the dichotomy, additional evidence continues to accumulate to support the existence of two different processing modes in speech perception." Pisoni does however acknowledge some exceptions to these observations.

Speech mode supporters do not have unquestioning support for their cause. Repp (1986) observes that "little has been learned from...(certain) studies about speech perception beyond the truism that perception within categories is not categorised...Equivalent information could have been obtained by using nonspeech stimuli."

This certainly seems to be true for some aspects of speech perception, for example, Fujisaki and Sekimoto (1975) were able to demonstrate perceptual overshoot (a common feature in speech perception) in both speech and non-speech conditions. Pols et al (1984), however, were unable to find a consistent overshoot effect. So it would appear that, as Repp (1986) agrees, one can specifically design experiments to exhibit categorical tendencies, but this does not mean they exist for all ways of performing such studies.

The view of psychophysics seems to be, in Repp's eyes, that one can demonstrate the same non-linearities with psychoacoustic stimuli as with speech if the experiment is designed in such a way as to do so. This contrasts somewhat sharply with the view of other workers at Haskins Laboratories (where categorical experiments were first performed); they are well represented by Mattingly and Liberman who state that "Perception of speech rests on a specialised mode, narrowly adapted for the efficient production and perception of phonetic structures.

Both sets of views tend towards the extreme, the truth lying probably somewhere between the two camps.

CHAPTER 2

Methods

2.1 Generation of Stimuli

All the synthetic speech stimuli used in the following experiments were generated, off-line, by an Apple IIe microcomputer using a software parallel-formant speech synthesizer. The synthesis program is a software implementation of the JSRU parallel-formant speech synthesizer as described in Rye and Holmes (1982) (which contains the source software). The original program was translated from Fortran to Pascal by D.J. Munden for use on the IBM PC and subsequently adapted and extended by the author for the Apple IIe microcomputer employing UCSD Pascal.

All the stimuli consist of four formants which are individually generated in a three-stage sequence. First of all, voiced and unvoiced excitations are mixed together (which in the present case is not used as all of the vowels used were totally voiced). The glottal pulse shape applied to the excitation generator is detailed below.

The second stage is a formant resonator with dynamic frequency control and bandwidth variation by rule (see Rye and Holmes, 1982). Finally, each formant is individually filtered with a spectrum weighting filter (Holmes, 1982) which substantially reduces any components outside the frequency range permitted to the associated formant. After the required number of formants have been produced they are combined in alternate polarity (i.e. formants are added in 180 degrees alternating phase).

To generate the synthetic vowel sounds a file (or list) of parameters are supplied during program execution. These stimulus

parameters were constructed from tables published in Holmes et al (1964), and Ainsworth (1974). The parameters supplied to the program are all sampled by the software at the same rate. This is fixed at 100 frames per second, i.e. every 10 ms.

The vowels /3/, /i/ and /a/, were chosen to give a wide representation of F1 - F2 space with a minimal sample of stimuli. Fig. 2.1 is taken from Ainsworth (1976) and illustrates the twelve vowels of English, in F1 - F2 space, spoken by a typical male speaker.

Fig. 2.2 shows the power spectra of the three vowels used. These were produced via a discrete Fourier transform (DFT) program. Each transform consists of 512 points represented on a linear frequency scale ranging from 0 to 5 kHz (abscissa). The ordinate shows relative level (expressed in dB) each value having a negative value with respect to the largest harmonic peak (which always has the value of 0 dB).

Fig. 2.2 further illustrates the reasoning behind the choice of vowel stimuli. /3/ was chosen as the first vowel for the pilot study (Chapter 3) and the main body of experimentation on frequency transition detection (Chapter 4), as it is the most central vowel in F1 - F2 space; it has relatively evenly and widely spaced formants. /a/ was used as it has a comparatively high F1 and low F2 frequency, and finally /i/ due to its low F1 and high F2. The exact significance of the above three choices will be discussed later in Chapter 4. The parameters of the synthetic vowel sounds are listed below in Table 2.1. They are as follows:-

Fig. 2.1

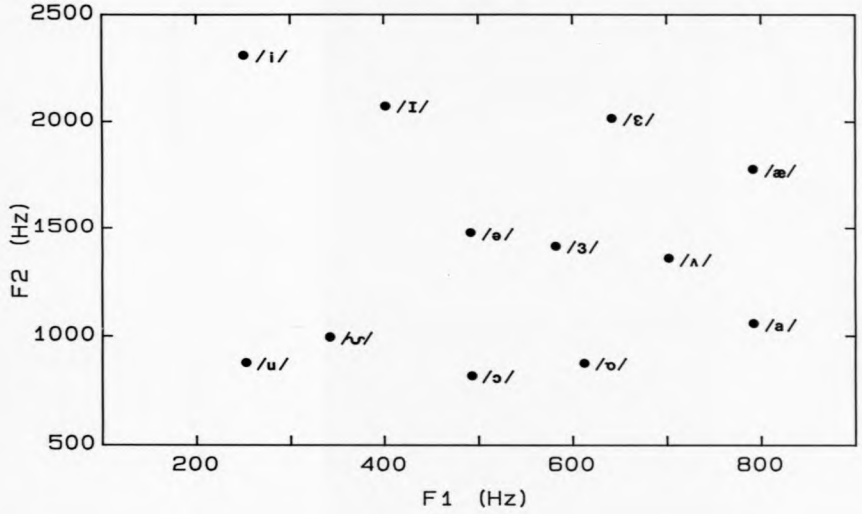

2500
2000
1500
1000
500
F2 (Hz)
200
400
600
800
F1 (Hz)
/i/
/I/
/ɛ/
/æ/
/e/
/3/
/ʌ/
/ɑ/
/ɑr/
/u/
/ɔ/
/o/

Fig. 2.2

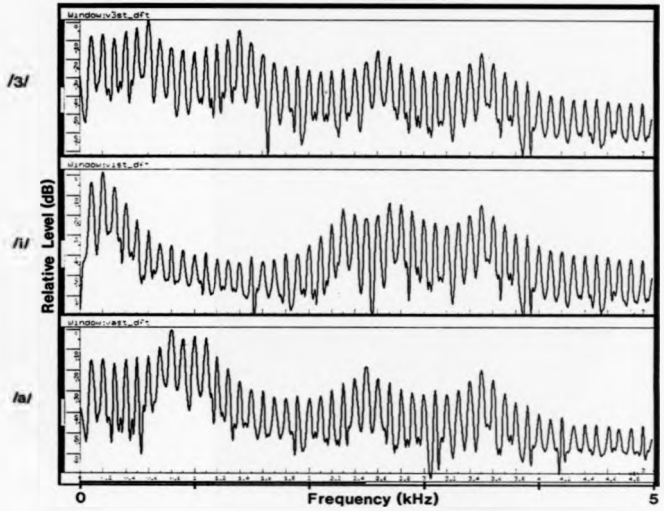


Table 2.1

	/3/	/i/	/a/
Frequency (Hz)			
F1	580	250	790
F2	1420	2320	1060
F3	2620	2740	2500
Amplitude (dB)			
F1	51	51	51
F2	45	33	49
F3	26	31	23

The above formant amplitudes are given relative to an an arbitrary reference; the absolute values are dependent on the overall stimulus level.

All the vowels had a fixed F4 = 3500 Hz, and a fixed F0 = 100 Hz. 100 Hz was chosen as the fundamental frequency as it is considered to be the average frequency of the male voice (Horst, 1982). The formant bandwidths were 80, 100 and 120 Hz for F1, F2 and F3 respectively. This was a somewhat arbitrary choice, as these values were maintained for different vowels despite variations in both F1 and F2 frequencies (as illustrated in Table 2.1). It was not thought, however, that fixed frequency formant bandwidths would have much effect on the following sets of experiments. Dunn (1961), in a paper on vowel formant bandwidth measurement, asserts that "...formant bandwidth values are not critical for the recognition of synthetic vowels." He goes on to say that exact values would probably give more natural vowel sounds. House (1960) looked at preferred bandwidths of vowels and found that a widening of formant bandwidths resulted in a decrease in 'naturalness'. Preferences depended on the vowel used, and the values used in the present study are on average slightly larger than those maximally preferred by House's subjects.

This approach, in retrospect, was possibly a little unwise. Horst (1982), for example, found that an increase in spectral slope (i.e. bandwidth decrease) of signals with a triangular spectral

envelope improved listeners frequency discrimination thresholds.

Each stimulus duration was 200 ms which is a fairly standard length for many psychoacoustical speech perceptual tasks of the nature reported below. All the synthetic vowels were generated using a Rosenberg glottal pulse shape in the synthesis program. This was employed because it is a good approximation to a human glottal pulse shape (Rosenberg, 1971); it produces reasonably "natural" sounding synthetic speech. Rosenberg (1971) used several different pulse shapes, in his preference experiments, in conjunction with varying combinations of opening and closing times (of the glottis). The pulse shape that bears his name (and the one used for the current experiments) is the one whose combinations of shape and relative opening and closing times was most preferred.

The isolated-formant stimuli were produced via the speech synthesizer; they are neither pure sinewaves nor noise bands, as employed by some workers, but a formant with glottal pulse excitation at F0. The resulting sounds are not particularly "speech-like", but serve as the most appropriate comparison stimulus where an analogous non-speech signal is required (for example, see Pols et al, 1984). Apart from synthetic vowel stimuli, sinewaves were used for some conditions; these were produced off-line, by a digital sinewave synthesis program.

All the stimuli (unless stated otherwise) were gated with a Hanning window to produce a 5 ms rise/fall time in order to remove any onset or offset clicks. The windowing was done off-line by multiplying the first and last 50 points of the digital waveform by the first and last 50 points of a 100 point Hanning window. (The sampling rate for all the digitised waveforms was 10 kHz, and therefore 50 digits represent 5 ms of analogue waveform at that digital-to-analogue conversion rate).

A Hanning window is an inverted DC-shifted cosine function with a maximum value of one. It was chosen in preference to the more popular Hamming window as its skirts move down to zero, whereas the skirts of a Hamming window are slightly raised on a pedestal. The Hanning window was also thought to be preferable to a linear ramp, as at durations of 5 ms the linear slope is possibly initially too steep to eliminate the percept of an initial or final click. The Hanning window function was constructed from formulae listed in Lynn (1982).

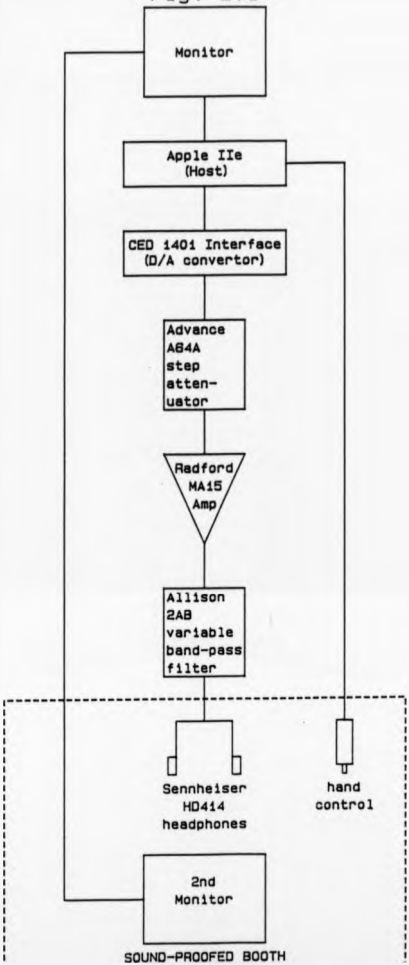
2.2 Apparatus and Procedure

2.2.1 Apparatus

A schematic diagram of the apparatus is shown in Fig. 2.3. The apparatus was controlled by computer software. At the start of an experiment, digitised stimuli were transferred from disc, via the host computer memory, to the user memory of the CED 1401 laboratory interface. After being converted to analogue signals by the CED 1401 interface, the stimuli were then presented binaurally through Sennheiser HD414 headphones to a subject seated in a sound-proofed booth. The digital-to-analogue output channel on the laboratory interface employed 12-bit data conversion which is standard resolution for a speech communication device. The signal was then fed in to an Advance A64A step attenuator, a Radford MA15 mono amplifier, and finally an Allison 2AB variable band-pass filter before being played through the headphones. The attenuator and filter were both passive devices, the filter being set to a passband of 100 - 5000 Hz. The filter's roll-off rate was approximately 36 dB per octave.

The sampling rate, as mentioned above, was 10 kHz; in order to avoid any distortion arising from aliasing during digital-to-analogue

Fig. 2.3


Monitor
Apple IIe
(Host)
CED 1401 Interface
(D/A convertor)
Advance
A84A
step
atten-
uator
Radford
MA15
Amp
Allison
2AB
variable
band-pass
filter
Sennheiser
HD414
headphones
hand
control
2nd
Monitor
SOUND-PROOFED BOOTH

conversion a high frequency cut-off at half the sampling rate was chosen.

2.2.2 Procedure

A two-alternative forced choice procedure was employed for all of the experiments. For each particular task thresholds were obtained at the 75 % correct level. Each subject responded to stimulus pairs presented at 6 different levels of difficulty in pseudorandom order. The main problem with a non-adaptive threshold estimation method such as this was how to choose a range of levels that are broad enough to encompass responses both below and above threshold, but also far enough away from each other to produce a graded scoring of responses. An interval ratio where an "all or nothing" situation exists in which listeners would score 100 % on one stimulus level and little better than chance on the next was as undesirable as one where the ratio was too small to encompass responses below and above threshold.

The criteria for the selection of six different levels of difficulty was therefore as follows:-

1) The easiest level should produce a score of (or close to) 100 % correct.
2) The hardest level should be adequately below threshold (75 % correct).
3) The intermediate levels should be narrow enough to produce a variety of scores such that the difference between two adjacent levels would never be near chance on the lower and near 100 % level on the upper of the pair.

4) The stimuli overall should be of sufficient difficulty to encourage the subject to concentrate (visual feedback was given), but not so hard that it became difficult to establish and maintain which perceptual criterion to use to distinguish differences between a test and a reference stimulus.

1) and 2) were essential for accurate and reliable threshold estimations; 3) was almost universally achieved although intersubject variability could be quite wide for a particular set of stimuli. 4) was less successfully achieved and possibly the most difficult to assess. Nevertheless, the method appears to be robust, and has produced fairly consistent and repeatable scores from the subjects who took part. An assessment of the reliability of the method appears in Chapter 5.4.

General psychophysical considerations suggested that a logarithmic stimulus scale in which equal divisions represent equal ratios of rate-of-change of frequency (for frequency transition detection experiments) or amplitude (for intensity transition detection) would be more appropriate than a linear scale. The interval ratio eventually chosen for frequency transition detection experiments, after some trial and error, was $\sqrt{2}$. On the whole this ratio worked very well. This resulted in widely varying ratios between adjacent terminal frequencies in experiments where rate-of-change of frequency was varied and duration held constant (Chapters 4 - 6). The ratios between adjacent terminal frequencies in this case was dependent on which formant, vowel and fixed duration was used (especially as the shortest transition duration of 10 ms with its highest thresholds produced much higher rates than a normal rate versus time trade-off). Relevant ranges are provided in each of the following experimental chapters. The choice of interval between

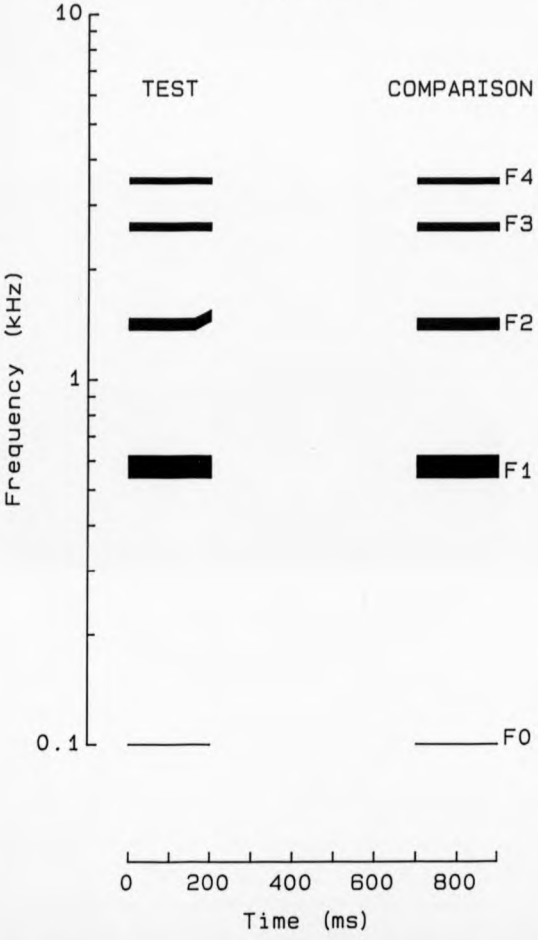
successive stimuli for the intensity transition detection experiments was slightly more complex, and is detailed in Chapter 7.2.

All newly generated stimuli were tested by the author first, and if the above criteria were satisfied they were tested on other subjects. If subjects found a series too difficult or too easy then the whole series would be shifted to match their ability at the task.

The subject was seated in front of a monitor which gave visual feedback of errors. Listeners were allowed as many presentations of the reference stimulus as they liked before the commencement of a trial. Each stimulus pair contained a test and a comparison (commonly a steady-state) stimulus, the task for the subject being to identify by means of a hand-held two-way switch which of the two intervals contained the test stimulus. (See Fig. 2.4 for a schematic representation of the task of frequency transition detection). The first stimulus of a subsequent pair was presented 500 ms after the response switch had been pressed.

Each subject was given several trial runs until consistent performance was achieved, i.e. until a plateau of correct responses was reached, but no attempt was made to overtrain. The order of stimulus and comparison was randomised for each pair. Each trial contained approximately 200 stimulus pairs and took, on average, around 10 minutes to complete (with stimulus lengths of 200 ms and an interstimulus interval of 500 ms). The interstimulus interval chosen was recommended as being long enough to avoid interference between successive stimuli, but also short enough for a comparison between stimuli to be made (Robinson and Watson, 1972). Three trials were added together for estimation of each individual threshold.

Fig. 2.4


10
1
0.1
Frequency (kHz)
TEST
COMPARISON
F4
F3
F2
F1
F0
0
200
400
600
800
Time (ms)

2.3 Subjects

A total of 10 subjects took part in the following experiments, though not all of them were used for every test. All of the subjects apart from one had normal audiograms; MLW had above normal loss in one ear at 4 kHz and he was therefore not used beyond the pilot study reported in the next chapter. The subjects were either postgraduate students or members of staff in the department. Postgraduate students who committed themselves to more than a few hours of experimentation were paid hourly for completed experiments.

Before commencing a set of experiments the nature of the task was explained to the subject concerned. Also a set of instructions, as listed in Appendix A, was displayed on the monitor screen in the sound-proofed booth.

2.4 Analysis of Data

As mentioned in section 2.2 above, thresholds were obtained at the 75 % correct level of performance. Raw data were fed into a program designed to fit a logistic function, that maximised the likelihood, to binary 'stimulus-response' data. The method, as reported in detail in Foster (1986), gives estimates for the mean (threshold) and the spread of the data. The standard deviation of the threshold was estimated by an incremental technique. A single threshold determination was derived from a total of 600 responses (100 at each of the 6 different levels of difficulty). The computer employed for threshold calculation was a CDC 7600 at the University of Manchester Regional Computer Centre, using software written by D.H. Foster.

Before the data could be statistically analysed, the levels of

difficulty (at an interval ratio of $\sqrt{2}$) had to be converted into levels separated by a linear number scale. The scores did not require any conversion as they were expressed as percentages. To convert from a ratio to a linear scaling the $\log_2$ of each level was obtained. (As the interval ratio was $\sqrt{2}$ the levels, after conversion, became equally spaced at intervals of 0.5).

Once the mean and spread were obtained then these values were reconverted to the original scaling by taking the antilog_2 of the mean. The threshold values could then be translated into meaningful units according to the performed task. Details relevant to specific tasks will appear in the results sections of the following chapters.

CHAPTER 3

Frequency Transition Detection: A Pilot Study

3.1 Abstract

The thresholds for frequency transition detection in the following experiments were obtained with the rising second formant of the vowel /3/ in both four- and isolated-formant conditions. By varying transition duration, fixed rates of frequency change were used. These were 4 and 8 Hz/ms for all subjects and in both conditions, and between 1 and 32 Hz/ms in doubling steps for one subject (in the four-formant condition). Contrast tests performed on pooled data showed that isolated-formants were easier to detect in the 4 Hz/ms case, but a reverse in detectability was found for the 8 Hz/ms data. When combined, there was no significant difference between four- and isolated-formant conditions; for rates-of-change between 1 and 32 Hz/ms a consistent increase in threshold with increasing rate-of-change of frequency was obtained (i.e. with decreasing duration of transition).

3.2 Methods

The formant transitions all occurred at the end of the stimulus in order to minimise pitch cues from steady-state segments. Tsumura et al (1973) proposed that if the last part of a stimulus is long enough to generate a definite pitch (i.e. > 100 ms) then the task is one of frequency discrimination and not frequency transition detection. For this reason all the initial experiments on frequency transition detection employed stimuli with transitions in final

position.

It was decided for this first set of experiments to fix the rate-of-change of frequency and, by so doing, to vary transition duration. Fixed rates of between 1 and 32 Hz/ms were employed. These are consistent with rates employed by other workers for this frequency range (Liberman et al 1956; Lehiste and Peterson, 1961; Holmgren, 1979; Pols and Schouten, 1982); consequently, to obtain thresholds for this set of transition rates, transition durations of between 4 and 102 ms were used. Typical durations of formant transitions lie in the range 10 - 100 ms (Nabelek and Hirsh, 1969; Tsumura et al, 1973). The total duration of each stimulus was 200 ms which is fairly standard for speech perception experiments (Pols, personal communication). The start frequency of the formant transitions was 1420 Hz in all cases (i.e. the F2 of /3/). The stimuli for this pilot study were not gated, i.e. their onsets and offsets were not windowed by a non-rectangular function. All the stimuli in subsequent experimental work were onset and offset gated, partly due to the findings of the present study.

The stimuli were presented at an overall sound level of 85 dB SPL. This proved to be an optimal and comfortable listening level, and ensured that all formant amplitudes were above threshold. Danaher et al (1973) found that at high sound levels (in their case 95 and 105 dB SPL) discrimination ability was reduced. As with all the experiments in this series the results represent data from partly trained subjects.

The specific instructions given to subjects before each experiment are detailed in Appendix A. The current methods section only refers to any differences from the basic paradigm (as presented in detail in Chapter 2). This is the case also for all following experimental chapters, though there is sometimes repetition of

material where it is felt that certain aspects of the design need re-emphasising (or the reader reminding of specific details).

3.3 Results

Table 3.1 Frequency transition detection thresholds (%)

		Transition rate			
		4 Hz/ms		8 Hz/ms	
		four-formant	isolated-formant	four-formant	isolated-formant
S U B J E C T	PC	5.31±0.2 (75.46)	4.68±0.18 (66.42)	5.88±0.21 (83.56)	6.2 ±0.21 (88.1)
	FM	4.74±0.28 (67.31)	5.89±0.27 (83.69)	6.06±0.3 (86.1)	6.84±0.34 (97.07)
	RCK	5.93±0.2 (84.18)	4.96±0.2 (70.49)	6.39±0.23 (90.78)	7.59±0.26 (107.84)
	RJB	5.97±0.25 (84.72)	5.69±0.23 (80.82)	6.56±0.21 (93.12)	7.14±0.34 (101.36)
	AT	4.99±0.15 (70.89)	4.26±0.18 (60.55)	7.54±0.21 (107.1)	6.94±0.19 (98.56)
	MLW	6.18±0.24 (87.79)	5.85±0.2 (83.13)	6.81±0.28 (96.65)	7.14±0.33 (101.41)
	JPW	6.01±0.21 (85.31)	5.2 ±0.18 (73.86)	6.86±0.22 (97.39)	7.0 ±0.25 (99.39)

Table 3.1 shows formant frequency transition detection thresholds for each subject in both four-formant and isolated-formant vowel conditions. The thresholds are expressed in terms of percentage of F2 start frequency. Standard deviations for each threshold are given for each value. The figures in parentheses represent the thresholds in terms of frequency (measured in Hertz). The F2 start frequency in all cases was 1420 Hz.

The figures in Table 3.1 show, as might be expected, that the thresholds for the 4 Hz/ms transitions appear to be lower in the isolated condition than in the four-formant condition. However, it can also be seen that this trend is reversed for the 8 Hz/ms transitions. In order to try and make more sense of this apparent paradox and also to ascertain which of the differences between the

thresholds in the four-formant and isolated-formant conditions are significant, statistical tests were carried on the data for both transition rates.

For a detailed description of the methodology of the particular type of statistical testing used, and the way they have been applied, see Appendix B. First of all, t-tests were employed on the columns of data in Table 3.1. Neither the 4 or 8 Hz/ms transition rates showed any significant difference (between isolated- and four-formant conditions) at $p<0.05$. The t-test was performed on frequency pairs expressed in terms of frequency as opposed to percentages. This was done because it was felt that this would emphasise any differences between pairs.

As the t-tests did not produce an acceptable level of significance a Wilcoxon matched-pairs signed-ranks test was then performed on the data. The Wilcoxon test uses information only about the direction of differences whereas the t-test uses information about the relative magnitude of the differences between scores, and hence is more powerful. Not surprisingly then, the Wilcoxon test also failed to show any significant difference between the four- and isolated-formant conditions for either of the transition rates tested.

Finally, contrast testing was employed on the data in an attempt to make use of the standard deviations provided by the "binfit" program. The methodology of contrast testing is discussed in detail in Appendix B. Both the t-test and Wilcoxon test were unable, by design, to make use of this potentially important information.

Using the data from the first two columns of Table 3.1, the formant transition detection thresholds for the four-formant vowel stimuli were found to be significantly higher than for the isolated second formant in the 4 Hz/ms condition ($p<0.001$, 1-tailed; $p<0.01$,

2-tailed). Conversely, for the 8 Hz/ms rate of frequency change (as represented in columns 3 and 4 of Table 3.1) the rising F2 of the vowel /3/ was harder to detect in the isolated than in the four-formant condition ($p<0.01$, 1 or 2-tailed). The latter result is the most surprising as it is in the opposite direction to that which was expected.

All the data were then taken together and a further test was performed to see if there were any differences between the isolated- and four-formant conditions regardless of transition rate. In this case there was no significant difference between the two conditions.

Table 3.2 Pooled data

	Transition rate			
	4 Hz/ms		8 Hz/ms	
	four-formant	isolated-formant	four-formant	isolated-formant
%	5.56±0.22	5.19±0.2	6.57±0.24	6.97±0.27
Hz	(79.02)	(73.67)	(93.25)	(98.95)

The pooled data thresholds, as shown in Table 3.2, represent values obtained by averaging the individual thresholds and standard deviations from the "binfit" program. The two populations (i.e. four- and isolated-formant) actually overlap in both 4 and 8 Hz/ms conditions. This can be demonstrated by adding and subtracting the standard deviations from each of the pooled threshold values; for the 4 Hz/ms transitions this produces values in the range of 5.34 - 5.78 % and 4.99 - 5.30 % for the four- and isolated-formant conditions respectively. The ranges for 8 Hz/ms are 6.33 - 6.81 % and 6.7 - 7.24 %.

Table 3.3 Individual data for four-formant condition (subject = PC)

	Transition rate					
	1 Hz/ms	2 Hz/ms	4 Hz/ms	8 Hz/ms	16 Hz/ms	32 Hz/ms
%	2.58±0.13	4.46±0.11	5.31±0.2	5.88±0.21	7.29±0.25	11.14±0.52
Hz	(36.7)	(63.37)	(75.46)	(83.56)	(103.54)	(158.24)

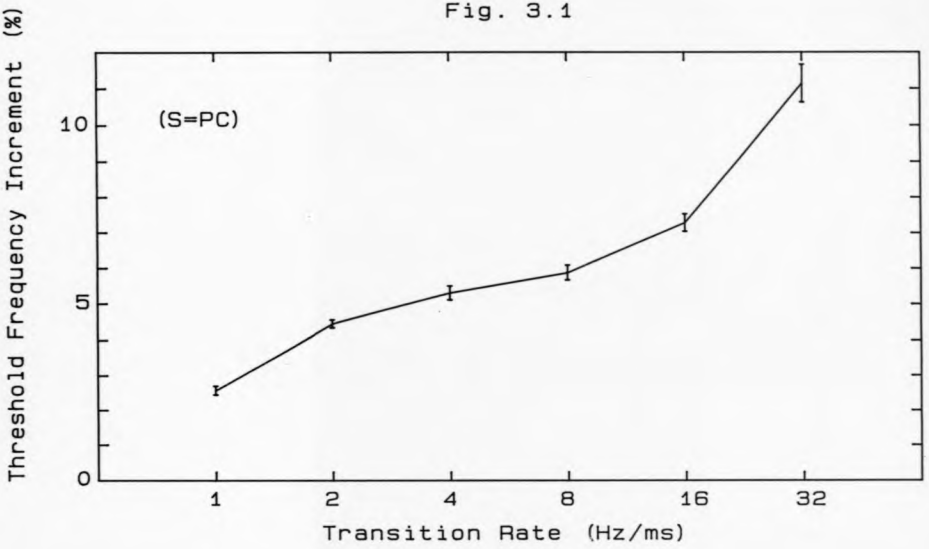
Table 3.3 and Fig. 3.1 both illustrate frequency transition detection thresholds obtained on one subject in the four-formant condition. The thresholds show a consistent tendency to increase as the rate-of-change of formant frequency also increases. The possible reasons for this trend are discussed in the following section.

3.4 Discussion

The reasoning behind the use of isolated-formants as a non-speech analogue was, hopefully, to demonstrate a difference in frequency transition detection thresholds between a speech sound and a comparable non-speech sound. On a masking model one would expect isolated-formant thresholds to be consistently lower. On a speech feature detection model the opposite prediction could be made. (One would predict the latter on the assumption that such a model works most efficiently with actual speech, however brief, rather than an isolated-formant which is, in effect, only a 'spectral segment' of a speech sound). Unfortunately, the predicted difference was in the opposite direction in the 8 Hz/ms case, though the contrast test combining both 4 and 8 Hz/ms data was not significant at any acceptable level. It would seem that further experimentation with different rates is necessary to see if any significant differences can be found elsewhere.

Two particular problems with experimental design emerged during the execution of the above experiments. Firstly, the data may well

Fig. 3.1


Threshold Frequency Increment (%)
(S=PC)
10
5
0
1
2
4
8
16
32
Transition Rate (Hz/ms)

have been corrupted by temporal factors inherent in the design of the experiment. If we look at the range of transition durations used to obtain thresholds at 1 Hz/ms rate-of-change of frequency, and compare them with those used in the 32 Hz/ms experiment we see a range of 32 to 102 ms in the former and 4 to 13 ms in the latter case. In maintaining a constant rate of frequency change the author has introduced an alternative variable into the experiment. This will in all probability have some effect on the magnitude of the observed thresholds.

One can see from Table 3.3 and Fig. 3.1 that thresholds are much lower for smaller rates-of-change of frequency. This finding is consistent with Sergeant and Harris (1962) who found that thresholds for detection of glissandi decreased with increasing duration over a much wider range of transition durations (75 ms - 10 s). Although this trend is a real one, the temporal cues assisting judgment at a slower rate-of-change cannot be ignored. At slower rates-of-change the listener has much more time to perceive a transition taking place even though the frequency excursion in question may be the same as at a shorter duration transition with a faster rate. Just how much difference in duration there is between the extremes has been indicated in the previous paragraph.

The literature for frequency transition detection of pure tones (as partially reviewed in Chapter 1.4) indicates a far greater sensitivity, although one obviously has to take rate-of-change of frequency and frequency excursion into account. It is therefore more than probable that a frequency change occurring over a longer period will be easier to detect than the same amount of change with a more rapid transition rate.

In order to overcome this problem different rates-of-change of frequency were employed in subsequent experiments reported in later

chapters, in order to keep the duration of the transition constant. One unfortunate side effect of this approach, however, is that there may also be some effect of transition rate on formant transition detection thresholds despite the fact that many other workers feel that rate-of-change of formant frequency has little or no effect on phoneme perception. This has always been the problem with studies in frequency modulation, as it is very difficult to isolate one variable for observation whilst keeping all other factors constant.

One particular study conducted by Holmgren (1979) specifically tried to establish whether formant frequency target or rate-of-change of formant frequency provides more information for the perception of vowel identity. The stimuli used were ambiguous insofar as they were designed to produce the percept of one vowel, if target frequency was more important, and a different vowel if rate of formant frequency change was the main source of information. Holmgren's results were not clear cut, but there was a tendency for subjects to favour target frequency. This is one of the few studies that looks favourably on rate-of-change of frequency as an important factor in speech perception in as much as the author does not discount the possibility that it can make a contribution to the identification of a vowel.

Gay (1970), in a study designed to ascertain which acoustic cues are used for the identification of certain American English diphthongs, made a number of interesting findings. Amongst these Gay found that rate of formant frequency change was primarily responsible for separating the percept of /Oi/ from /ai/. He also states that the diphthongs ".../Oi,ai,au/ are characterised primarily by an invariant rate of formant frequency change." So this can be taken as further evidence to support the role of rate-of-change of formant frequency in phoneme identification, albeit a minor one in comparison to extent of frequency change or transition duration.

The second major problem encountered in the above experiments was that of gating of stimuli. Onset gating was not necessary as the speech synthesis software ramped onsets automatically. Offset gating, however, has proved to be quite problematic. Despite the fact that there is an exact number of glottal pulses in every 200 ms stimulus (i.e. there is no possibility of being in the middle of an excitation at the end of a stimulus) it is not possible to predict the effects of differing rates-of-change of frequency and variable frequency targets. When there is no transition in the waveform and as long as there are a whole number of glottal pulses contained in it then it should end on or very close to a zero crossing. With different durations and extents of transitions occurring at the end of a 200 ms stimulus, a zero crossing cannot be predicted or guaranteed. Consequently, there may well be steps in the waveform at the end of some of the stimuli, which may aid or confound responses and yet not be easily perceivable. Therefore, every stimulus (regardless of necessity), was both on and offset gated for all subsequent experiments. No subject reported the perception of a click at the offsets of any of the vowels, but with subsequent sinewave experiments a click was perceived if an offset was not gated, regardless of whether the stimulus finished on a zero crossing.

One reason that the differences between the four-formant and isolated-formant conditions were not significant could be due to the fact that the proximity of the formants in the /3/ phoneme was such that very little masking took place in the four-formant condition. In the design of the pilot study care was taken so that formants did not overlap during a transition, i.e. that the F2 of /3/ did not cross or overlap F3. This was done in order to minimise formant interaction in an attempt to produce reference data for later

experiments where the converse was sought. The vowel /3/ consisted of an F1 of 580 Hz, an initial F2 of 1420 Hz and an F3 of 2620 Hz, with bandwidths of 80, 100 and 120 Hz respectively. Therefore, at least during the steady-state portions of the stimuli, none of the first three formants are within a critical band of each other. (See critical band values given in Scharf, 1970; and more recently Sekey and Hanson, 1984).

One would expect differences between the two conditions to be more marked where F1 is much closer to the initial frequency of F2 (as in the phoneme /a/), or where the final F2 is closer to or moving across F3 (which would happen with the phoneme /i/). In the following chapter the phonemes /a/ and /i/ are used in order to observe possible effects of upward masking of F1 on F2, and also the change in detectability of a transition due to a transition in F2 coming within a critical band of F3 (and causing a sudden increase in amplitude of the harmonics of the fundamental in that frequency region). One would expect the former case to make a transition less detectable and the latter to make it more so.

In Chapter 5 further controls in the form of pure tones are employed. The reasons for this are twofold. Firstly, this will indicate (from the thresholds the pure tones produce) if the overall experimental method is sensitive enough. There is a certain amount of data from other workers investigating frequency transition detection in the non-speech domain and even more data on frequency discrimination, and so a comparison should be easy to make.

Secondly, listeners in the above experiments found the isolated-formant sounds to be quite 'speech-like' and consequently may well have produced qualitatively different thresholds than if the non-speech analogue used had been a noise band, for example. This depends on whether the assertion that humans use a different mode of

perception for speech than for all other sounds is true or not. There has been a lot of support for this point of view from many workers involved in speech perception and recognition, so the latter assertion cannot be discounted (Liberman et al 1967; Mattingly et al 1971). This particular issue is dealt with in more detail in a literature review in Chapter 1 above, and also experimentally in Chapter 6.

The thresholds obtained for the above experiments could be seen to be slightly on the large side when compared with data for pure tones. The above refinements to the design of the experiments were designed to produce more accurate measures of formant frequency transition detection thresholds. Danaher et al (1973), in a formant transition discrimination experiment, obtained thresholds of 50 - 80 Hz for the phonemes /a/ and /i/ (50 % correct level), but as the rate-of-change of frequency is not known in this case it is hard to make any real comparison. Nevertheless, 50 - 80 Hz is within the range of the present results.

Horst (1982) appears to have obtained some of the lowest thresholds so far. As reported above, he achieved sweep detection thresholds of between 0.4 and 2.0 %; but these values must be viewed in context. Although his stimuli were typically in the F2 frequency region, they were band-passed filtered pulse trains, with rectangular sweeps, ranging from 23 to 99 ms. Horst claimed that the introduction of jitter into the pulse train, due to smoothing of its long-term spectrum, gave the stimulus a more natural speech-like quality. Nevertheless, he did not report any of the subjects perceiving the pulse trains as speech.

No feasible explanation can be offered at present for the reversal of trend between thresholds in the four-formant and isolated-formant conditions for rates-of-change of frequency of 4 and

8 Hz/ms. Care must be taken in the interpretation of the results of the statistical tests. Out of the three different methods used only one of them managed to produce results of any acceptable level of significance. The differences that were demonstrated, by contrast analysis, were conflicting at best.

One may conclude from the above study that investigating frequency transition detection, in conjunction with the above methods, with rate-of-change of frequency as the independent variable produces contradictory results. However, a repeat of the experiments, with some tightening up of the methods, could well produce less conflicting data. The main benefits of the present study have been in testing and revising the above experimental method; the success of this can be measured in the results obtained in the following four chapters.

CHAPTER 4

Frequency Transition Detection using Synthetic Vowel Sounds

4.1 Introduction

The work discussed in the present chapter was reported in a preliminary form in Cosgrove and Wilson (1986). The current study presents this work in full and includes some previously unpublished results.

In Chapter 3 it was found that frequency transition detection thresholds obtained from the rising F2 of /3/ produced conflicting data if transition rate was held constant and duration was varied. Also, problems arising from possible temporal interactions arose. In a bid to avoid these pitfalls the experiments of the present chapter were carried out. The duration of transition was held constant and a varying rate-of-change of frequency was employed to obtain the thresholds.

4.2 Methods

4.2.1 Choice of Stimuli

The vowels /3/, /a/ and /i/ were used for the present study; the parameters of these vowels were previously supplied in Table 2.1. /3/ as mentioned above was chosen for its relatively evenly and widely spaced formants (on a critical band scale). Using the phoneme tables in Ainsworth (1974) and Fig. 2.1 (which illustrates the twelve vowel phonemes of English in their F1 - F2 space), it can be seen that /3/ has mid-range frequencies for both F1 and F2 for a vowel

phoneme. From Fig. 2.1 it can also be seen that /i/ has the lowest of the vowel F1 and highest of the F2 frequencies. Finally, the vowel /a/ has the highest of the F1 and less than average of the F2 frequencies.

The aims and intentions behind using /a/ and /i/ alongside /3/, for frequency transition detection experiments, were briefly outlined in the discussion section of the previous chapter. Specifically, where detection of a rising second formant is concerned, /i/ was chosen because its F2 and F3 are in such close proximity, and /a/ because of the nearness of its F1 and F2. With /3/ as a control vowel it was hoped that the effects of neighbouring formants could be observed on the frequency transition detection thresholds of the second and first formants of the vowels /i/ and /a/. (See Fig. 2.2 for an illustration of the above explanations).

Thresholds of both F1 and F2 transitions were obtained for all of the three vowels, though in the case of isolated-formants just the rising F2's of /3/ and /i/ were tested. The small number of isolated-formant conditions was partly mediated by a need to keep the stimulus set to a manageable minimum. Also, the two conditions used were adequately able to demonstrate the effect of analogous but less complex stimuli on the task of frequency transition detection.

4.2.2 Other Stimulus Parameters

Transition durations of 10, 20, 40 and 80 ms were chosen to be typical of the range of frequency transitions in speech (Nabelek and Hirsh, 1969; Tsumura et al, 1973). Within each fixed duration condition rates-of-change of formant frequency were varied. Rates of between 0.175 and 45.2 Hz/ms were used for F2 transitions, and between 0.0625 and 64 Hz/ms for those of F1. One reason for such

wide ranges is that rate (expressed in Hertz as opposed to octaves per millisecond) is dependent on test frequency, which differed widely between vowels and formants (cf. frequency resolution over a wide band of frequencies). However, the main reason was the four fixed rates of transition duration, the shortest of which needed rates at the top end of the above ranges.

As stated in Chapter 2, the stimuli were presented at six different levels of difficulty, the ratio between adjacent levels being set at $\sqrt{2}$. To give a typical example, for the F2 of /3/ (rising or falling) with a fixed duration of 80 ms, listeners were presented with rates of 0.175, 0.25, 0.35, 0.5, 0.7125 and 1 Hz/ms to obtain threshold. (There was a small amount of rounding necessary in order to produce integer amounts of frequency change as parameter values for the synthesis program).

The formant transitions all occurred at the end of the stimulus (as in the previous chapter) in order to minimise the pitch cues from steady-state segments. Except where otherwise stated, all transitions were upward. As stated in Chapter 2, all the stimuli were of 200 ms duration. The stimuli were presented at an overall sound level of 85 dB SPL. This meant that the same formant from different vowels would be at different sound levels (see Table 2.1); in natural speech the formants of different vowels have different intensities relative to each other anyway. Keeping the overall intensities of the three vowels equal is arguably a more natural approach than equalising particular formants.

4.3 Analysis of Data

As in the previous chapter formant frequency transition detection thresholds were obtained at the 75 % correct level.

Contrast analysis was then performed on both individual and pooled data. Comparisons were made between the detectability of four-formant and isolated-formant conditions, and between the first and second formants of a particular vowel. In cases where a subject had completed trials on the same formant for two different vowels a contrast analysis on the differences between the two was performed.

4.4 Results 1

As expected, all conditions show a decreasing threshold with increasing transition duration with greatest changes occurring between 10 and 20 ms (Figs 4.1 - 4.33). All the thresholds were expressed in terms of percentage frequency change (i.e. Δf / f x 100), as extent of change divided by a reference is a common and accepted method of representing psychophysical thresholds. It was not possible to illustrate the results from different stimulus conditions using the same y-scale throughout. Where required, for example, when the same stimulus conditions are being represented between several figures, a particular scale is maintained.

Figs 4.1 - 4.17 are all individual results which need not be considered separately as they are all represented in the graphs of pooled data thresholds in Figs 4.18 - 4.22; they are included to show the degree of consistency between subjects.

Fig. 4.18 shows transition detection thresholds of the F2 of /3/ under isolated- and four-formant conditions. Predictably, the F2 transition was slightly but significantly easier to detect when isolated than within the four-formant context ($p<0.05$, 1-tailed). The differences for /i/ were more marked than for /3/ ($p<0.001$, 1 or 2-tailed; see Fig. 4.19), possibly due to a downward masking effect of F3 upon F2 (see Fig. 4.36). It is assumed, from our knowledge of

Key to Chapter 4 Figures

F1 = first formant

F2 = second formant

FF = four formant

IF = isolated formant

U = upward transition

D = downward transition

N = number of subjects

S = initials of subject

(for individual data)

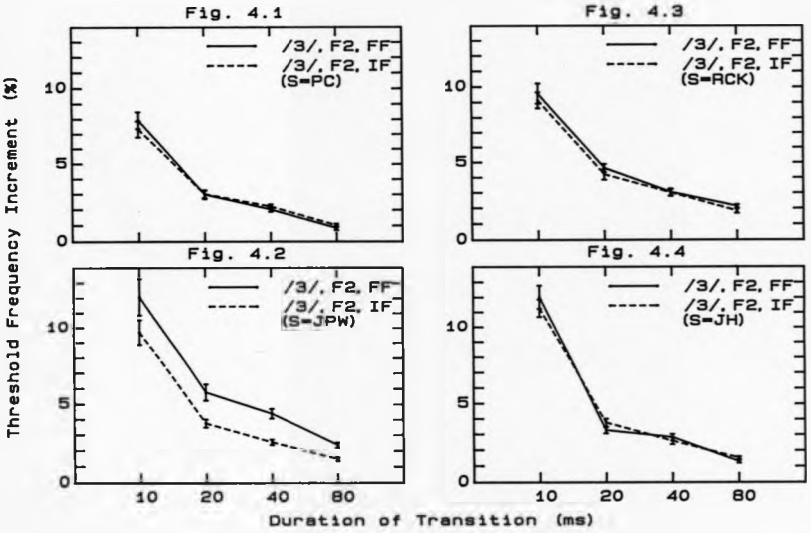


Fig. 4.1
/3/, F2, FF
/3/, F2, IF
(S=PC)
Fig. 4.3
/3/, F2, FF
/3/, F2, IF
(S=RCK)
Fig. 4.2
/3/, F2, FF
/3/, F2, IF
(S=JPW)
Fig. 4.4
/3/, F2, FF
/3/, F2, IF
(S=JH)
Threshold Frequency Increment (%)
0
5
10
10
20
40
80
Duration of Transition (ms)

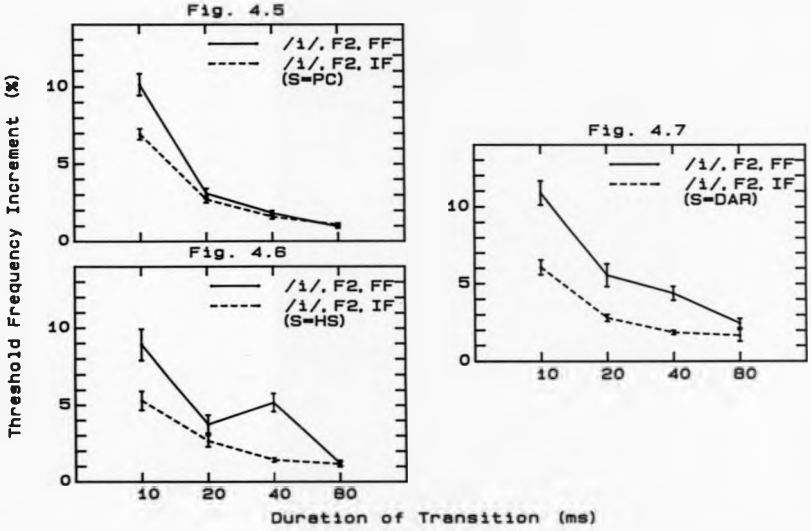


Fig. 4.5
/i/, F2, FF
/i/, F2, IF
(S=PC)
Fig. 4.6
/i/, F2, FF
/i/, F2, IF
(S=HS)
Fig. 4.7
/i/, F2, FF
/i/, F2, IF
(S=DAR)
Threshold Frequency Increment (%)
0
5
10
10
20
40
80
Duration of Transition (ms)

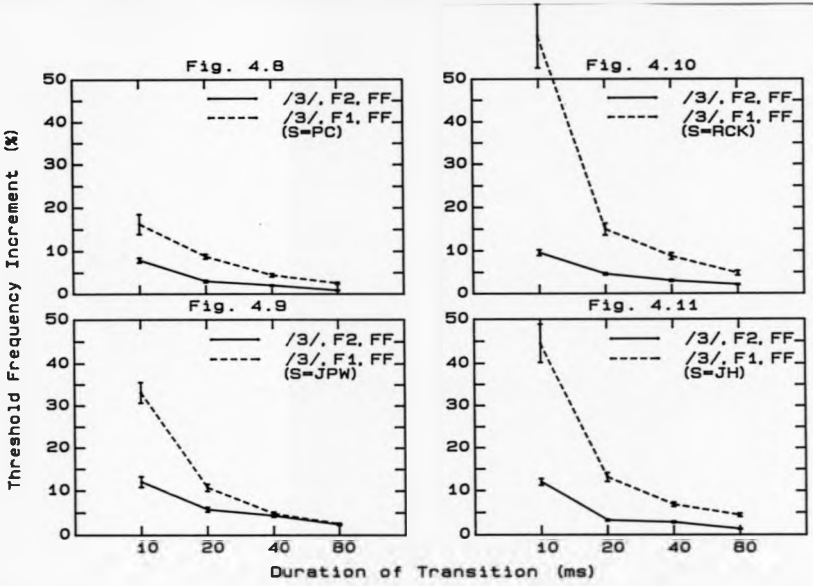


Fig. 4.8
/3/, F2, FF
/3/, F1, FF
(S=PC)
Fig. 4.10
/3/, F2, FF
/3/, F1, FF
(S=RCK)
Fig. 4.9
/3/, F2, FF
/3/, F1, FF
(S=JPW)
Fig. 4.11
/3/, F2, FF
/3/, F1, FF
(S=JH)
Threshold Frequency Increment (%)
0
10
20
30
40
50
10
20
40
80
Duration of Transition (ms)

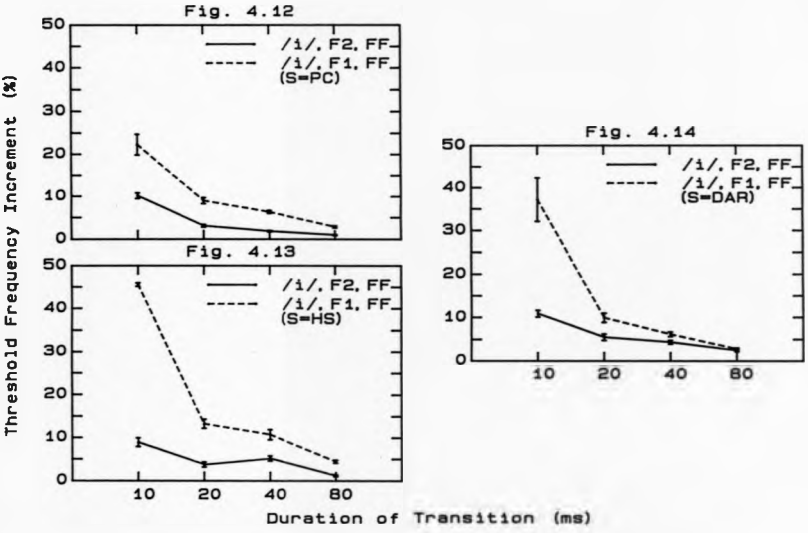


Fig. 4.12
/i/, F2, FF
/i/, F1, FF
(S=PC)
Fig. 4.13
/i/, F2, FF
/i/, F1, FF
(S=HS)
Fig. 4.14
/i/, F2, FF
/i/, F1, FF
(S=DAR)
Threshold Frequency Increment (%)
Duration of Transition (ms)
0
10
20
30
40
50
10
20
40
80

Fig. 4.15

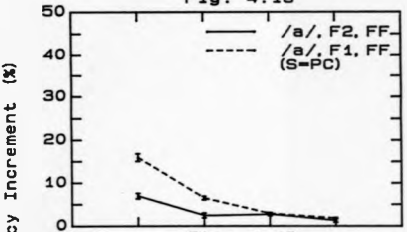
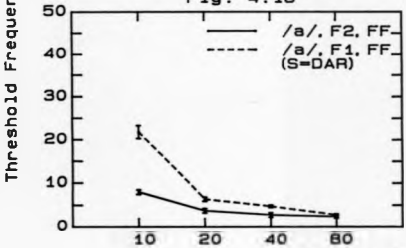


Fig. 4.16



Threshold Frequency Increment (%)

Duration of Transition (ms)

Fig. 4.17

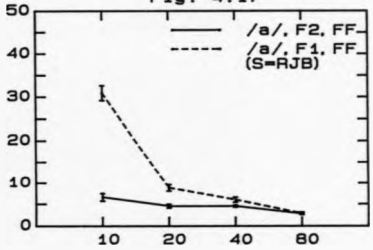


Fig. 4.18

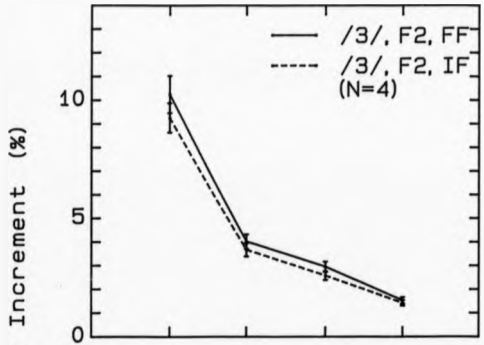
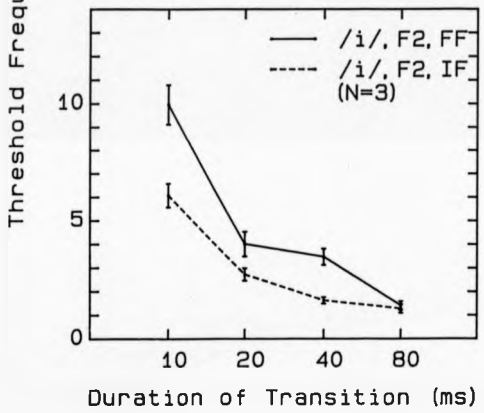


Fig. 4.19



how masking works (see Chapter 1.2), that when a dynamic frequency component (or set of components in the case of formants) moves to within a critical band of an adjacent static component, the detectability of the former will decrease. At detection threshold F2 and F3 were within one critical band (Scharf, 1970) for /i/ but much more widely spaced for /3/.

The isolated-formant, as mentioned in Chapter 2.1 is a formant with glottal pulse excitation at F0; this serves as the most appropriate comparison for the possible influence of other formants.

The transition thresholds for F1 detection were compared with those for F2 in all the three vowels tested. The individual comparisons can be seen in Figs 4.8 - 4.11 for /3/, Figs 4.12 - 4.14 for /i/ and Figs 4.15 - 4.17 for /a/. The pooled data are represented in Figs 4.20 - 4.22. From the pooled thresholds one can see that the differences are quite clear, requiring about twice as much change in F1 for detection. For /3/, /i/ and /a/ the differences were significant ($p<0.001$, 1 or 2-tailed) that the F2 transition was easier to detect than the F1.

Because of time constraints individual subjects tended to be tested on one particular vowel. With different subjects direct comparison between vowels becomes impossible. However, one subject (PC) performed all of the experiments in order that some direct comparisons could be made. Also, another subject (DAR) completed trials on all of the (rising formant) experimental conditions for the vowels /i/ and /a/.

In Fig. 4.23 the detection thresholds of the F2's of /3/ and /i/ are plotted. For one subject it was found that the F2 transitions of /3/ were easier to detect than those of the vowel /i/ ($p<0.05$, 1 or 2-tailed). This was in line with the pooled results shown in Figs 4.18 and 4.19, and individually in Figs 4.1 and 4.5 (for the same

Fig. 4.20

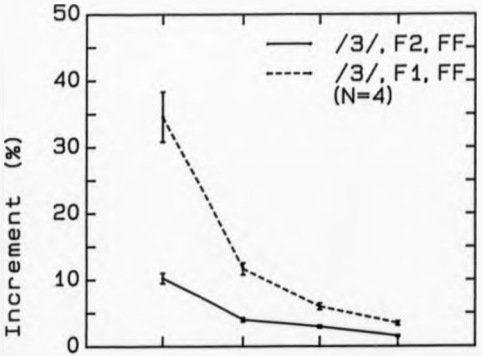


Fig. 4.21

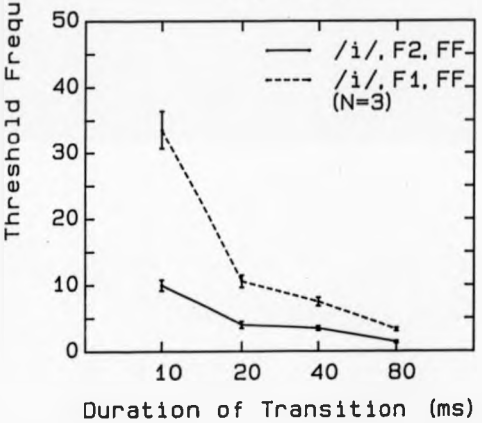


Fig. 4.22

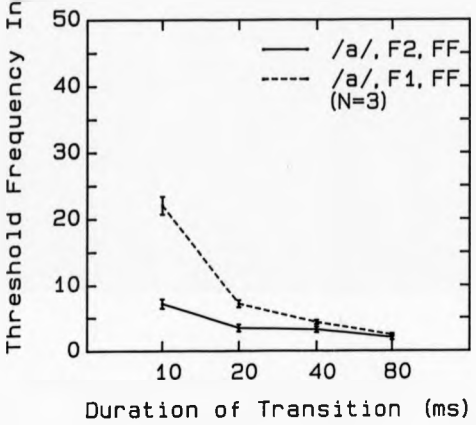

Threshold Frequency Increment (%)
50
40
30
20
10
0
/a/, F2, FF
/a/, F1, FF
(N=3)
10
20
40
80
Duration of Transition (ms)

Fig. 4.23

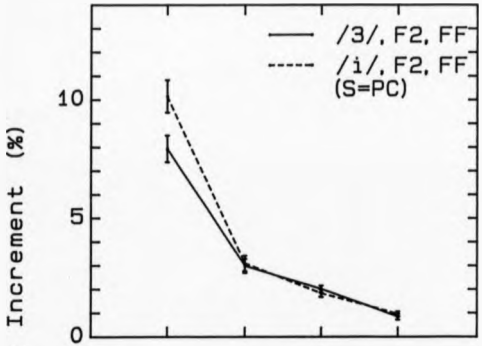
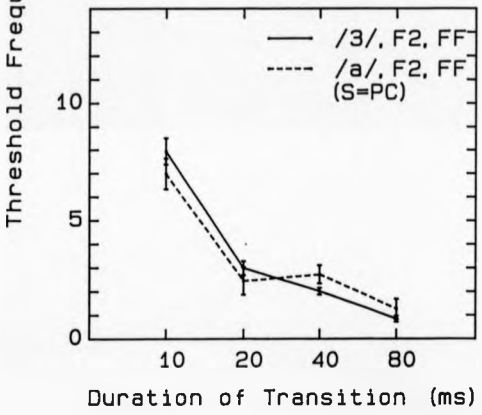


Fig. 4.24



subject). In Figs 4.5 and 4.19 the F3 of /i/ appeared to interfere with the detection thresholds of F2, whereas for /3/ this does not seem to be the case, due presumably to a wider and more even formant spacing (see Fig. 2.2).

Fig. 4.24 illustrates a comparison between the F2 transitions of /3/ and those of /a/ for one subject. No significant difference was found between the two stimulus conditions. In both cases there was no F3 to interfere with a rising F2 as there was for /i/. One would not expect any large difference between the two conditions due to frequency effects. The F2 frequencies of /3/ and /a/ used in the present experiments were 1420 Hz and 1160 Hz respectively. Moore (1974), Wier et al (1977) and others have illustrated a change in discriminative ability for frequency changes over wide ranges of test frequency. There is no indication however, that in the F2 region of /3/ and /i/ a difference in reference frequency of 260 Hz (F2 /3/ - F2 /a/) would have any particularly marked effect.

Figs 4.25 and 4.26 show a comparison between the F2 detection thresholds of /i/ and /a/ for two subjects. The result for one subject (Fig. 4.25) was significant ($p<0.05$, 1-tailed) that the rising F2 of /a/ is easier to detect than the F2 of /i/. The other individual comparison between these two conditions (see Fig. 4.26) is more emphatic ($p<0.001$, 1 or 2-tailed). These results could also be due to a possible masking effect of the F3 upon the F2 for the vowel /i/ as was previously suggested in the comparison between the F2's of /3/ and /i/ (Fig. 4.23).

Similar individual comparisons were made between the frequency transition detection thresholds of rising first formants of different vowels as for second formants. Fig. 4.27 shows the thresholds for the F1's of /3/ and /i/ for one subject; the result is significant ($p<0.01$, 1-tailed; $p<0.05$, 2-tailed) that the F1 of /3/ is easier

Fig. 4.25

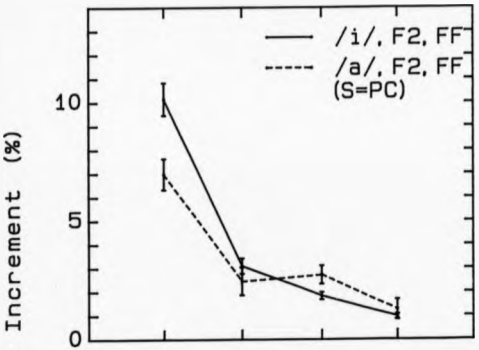
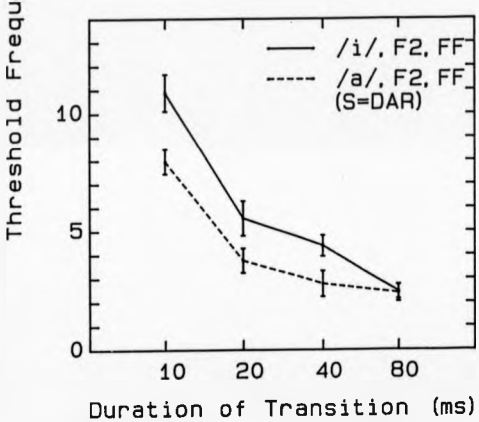


Fig. 4.26



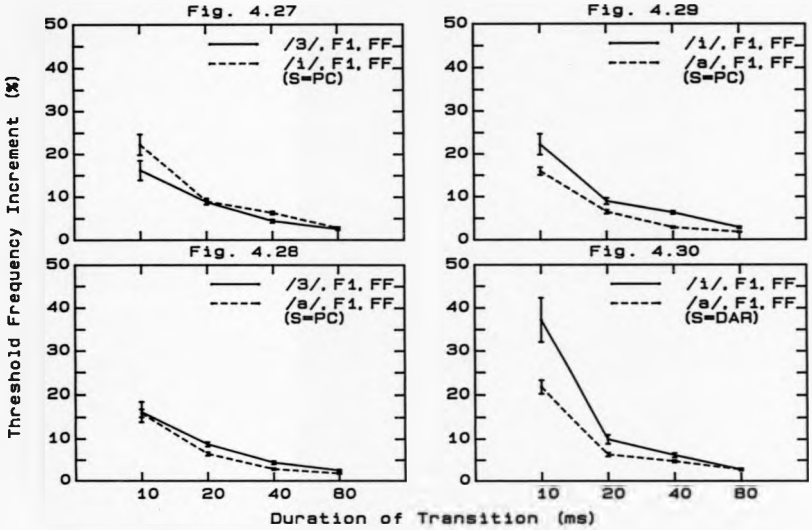


Fig. 4.27
/3/, F1, FF
/i/, F1, FF
(S=PC)
Fig. 4.29
/i/, F1, FF
/a/, F1, FF
(S=PC)
Fig. 4.28
/3/, F1, FF
/a/, F1, FF
(S=PC)
Fig. 4.30
/i/, F1, FF
/a/, F1, FF
(S=DAR)
Threshold Frequency Increment (%)
0
10
20
30
40
50
10
20
40
80
Duration of Transition (ms)

than that of /i/ to detect. Fig. 4.28 illustrates an individual comparison between the F1's of /3/ and /a/; /a/ proved to be easier to detect than /3/ ($p<0.05$, 1-tailed). This is a slightly odd result as one would have perhaps expected the F1 of /3/ to be easier to detect due to a possible downward masking effect of the F2 upon the F1 of /a/. No explanation can be offered for the result at present, though the difference between the two conditions is at the least of the acceptable levels of significance. Figs 4.29 and 4.30 show individual comparisons between the F1's of /i/ and /a/. For both subjects the result is highly significant that the rising F1 of /a/ is easier to detect than that of /i/ ($p<0.001$, 1 or 2-tailed). An explanation of this result is offered below in section 4.7 where a model and method for normalising data from different formant and vowel conditions is discussed.

One subject (PC) completed trials on both upward and downward transitions of all three isolated vowels. Gardner and Wilson (1979) found that at 1 kHz the threshold frequency change for 62.5 ms pure-tone upsweeps was lower (0.85 %) than for downsweeps (1.5 %). Collins and Cullen (1984), using a noise masking technique, also found lower thresholds for upsweeps. This result also agrees with other work in this area, for example, Cullen and Collins (1982) and Lacerda and Moreira (1982). In the present study for the F2 of /3/ (Fig. 4.31), downsweeps had a slightly lower threshold for the three shorter durations, and slightly higher at 80 ms. Overall, however there was no significant difference. Interestingly, the mean for the values (40 and 80 ms) around 62.5 ms and for up and downsweeps (1.42 %) agrees very closely with Gardner and Wilson's mean (1.18 %).

Fig. 4.32 compares upward and downward sweeps of the F2 of /i/. The result is significant that the rising F2 is easier to detect than the falling F2 ($p<0.05$, 1-tailed). This is a slightly contrary

Fig. 4.31

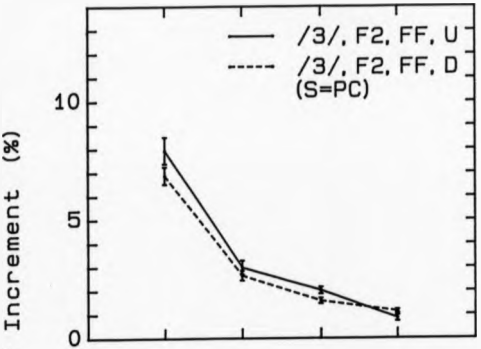
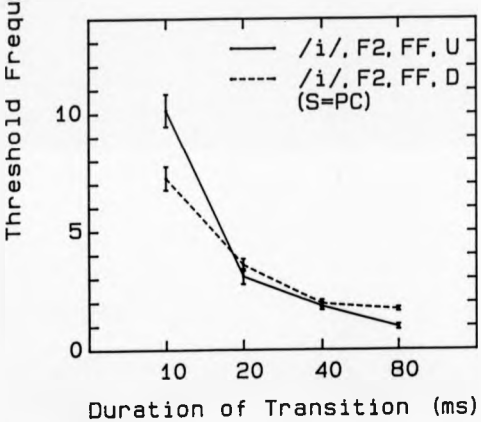


Fig. 4.32



Threshold Frequency Increment (%)

result especially as the F3 of /i/ has seemingly had an effect on the rising F2 in the comparisons between the F2's of /i/ and both /3/ and /a/ (see Figs 4.23, 4.25 and 4.26). The significance is low in this case, and perhaps a slightly greater sensitivity to upsweeps (Gardner and Wilson, 1979; Collins and Cullen, 1984) than for downsweeps may have had some effect on the result.

Finally, Fig. 4.33 shows a comparison between the rising and falling F2 of /a/. The result is strongly significant that the rising second formant is easier to detect than the falling F2 for the vowel /a/ ($p<0.001$, 1 or 2-tailed). This result agrees with the author's prediction. When the F2 of /a/ is falling it moves towards an F1 which is only 180 Hz away from it (when F1 and F2 bandwidths are accounted for). According to Scharf (1970) the critical band at 1 kHz is 160 Hz, and at threshold for each of the durations tested the F2 had moved to within a critical band of F1. Therefore, it is reasonable to assume that a downward transition is harder to detect than one that is rising with the F2 of /a/ due to upward masking on the downward condition from the first formant. There is one anomaly in the data however, as at 10 ms (when one would expect the difference between upward and downward transitions to be at a maximum) the thresholds are closest together.

4.5 Discussion 1

4.5.1 Temporal Integration

In all the figures so far discussed the threshold values all appear to decrease consistently as transition duration increases, so the pooled data were replotted as threshold (%) x duration (ms) as a function of transition duration. These are shown in

Fig. 4.33

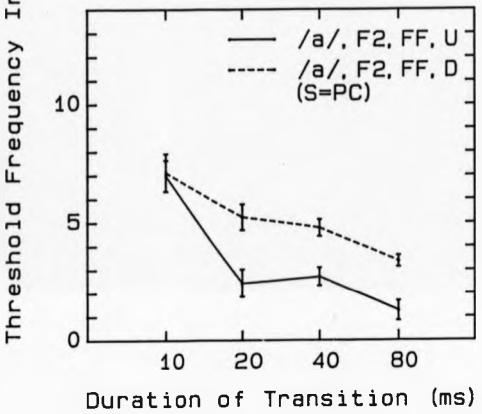

/a/, F2, FF, U
/a/, F2, FF, D
(S=PC)
Threshold Frequency Increment (%)
0
5
10
10
20
40
80
Duration of Transition (ms)

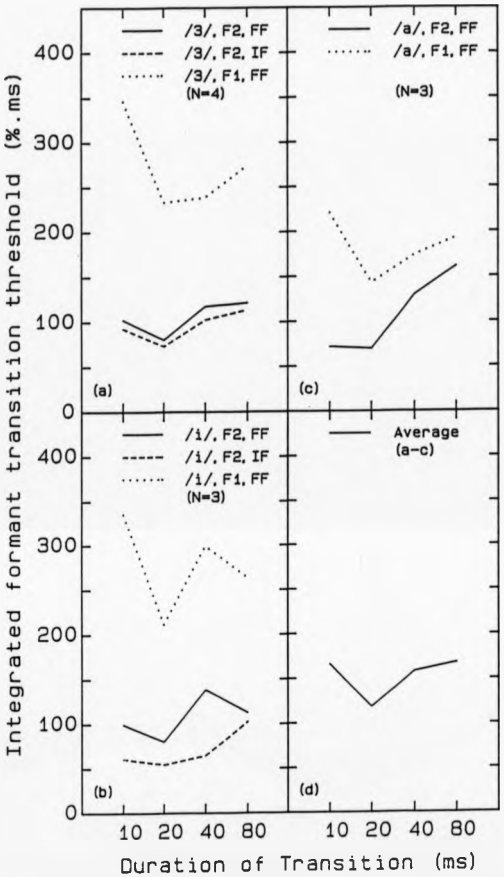
Fig. 4.34(a - c). On average (Fig. 4.34d) this relationship appears to be relatively independent of duration. The implication of this is that sensory integration is taking place, so that threshold formant frequency increment is inversely proportional to transition duration. (It should be noted that this is the inverse of rate-of-change detection where threshold would be proportional to transition duration).

Sensory integration (or temporal integration as it is sometimes called) in the above context refers to a mechanism of the auditory system whereby the ear seems to integrate the energy contained in a sound over a short period of time. The length of time involved is known as the integration time. Patterson (1986) defines the integration time as "...the sample duration over which the auditory system combines information when computing parameter values or average statistics". This time constant of integration varies with frequency (Moore, 1982) and the type of sound stimulus used (Patterson, 1986).

The integrative relationship demonstrated in Fig. 4.34 appears to hold over the majority of the range 10 - 80 ms, which covers (almost completely) the durations found in speech transitions. It appears likely, however, that it must break down for very short durations where the frequency excursion would become very large; and for longer durations where the integrating time constant of the auditory system would be exceeded. Stephens (1973a) neatly illustrates the above two assertions with a graph of mean detectability of equal energy tone bursts (1 kHz) plotted as a function of duration. Detectability steadily rises to a plateau at around 20 ms and begins to fall off a little above 200 ms.

This could explain why the greatest changes in frequency transition detection thresholds were observed between durations of 10

Fig. 4.34


/3/, F2, FF
/3/, F2, IF
/3/, F1, FF
(N=4)
/a/, F2, FF
/a/, F1, FF
(N=3)
/i/, F2, FF
/i/, F2, IF
/i/, F1, FF
(N=3)
Average
(a-c)
(a)
(b)
(c)
(d)
Integrated formant transition threshold (%.ms)
0
100
200
300
400
10 20 40 80
Duration of Transition (ms)

and 20 ms. If 10 ms is too short a time, within the experimental constraints, for proper sensory integration to occur then it is to be expected that the threshold values will increase for this duration. One practical implication of the sensory integration phenomenon is that related experiments need perhaps only be performed at a single duration within the 10 - 80 ms range, and the results could then be compared with other experiments where a different duration has been used.

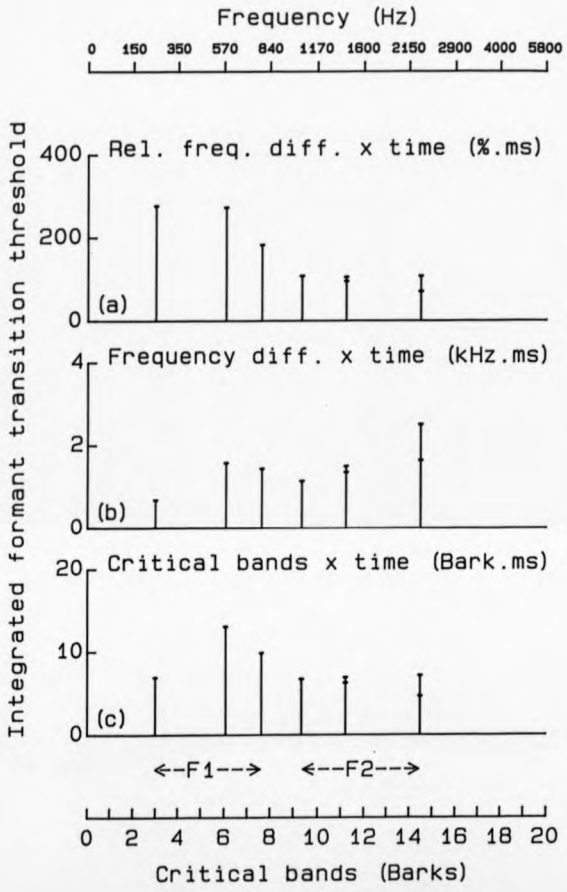
One interesting aspect of the integrated thresholds is that without exception the lowest threshold in each case occurs at the 20 ms duration of transition. The largest threshold, however, is fairly evenly distributed between the other three transition durations. One would ideally expect the data from Fig. 4.34 to describe part of a U-shaped function, the higher thresholds on either side occurring at durations where temporal integration is breaking down.

In an attempt to obtain further generalisation of the results, the mean values (threshold (%) x duration (ms)) from Fig. 4.34 were plotted against frequency on a critical band scale (Barks). The critical band scale was constructed from a table of values given in Scharf (1970). Although there have been subsequent revisions of the bark scale, for example, Sekey and Hanson (1984), Scharf's values are widely accepted and more universally adopted.

In Fig. 4.35a the threshold values (% x ms) appear to decrease systematically with critical band position of the formant concerned. If, however, threshold frequency increment x duration (Hz x ms) was plotted versus critical band position, the opposite trend was noted (Fig. 4.35b).

It was decided, therefore, to express the threshold frequency increments in critical bands (x duration) (see Fig. 4.35c). These

Fig. 4.35

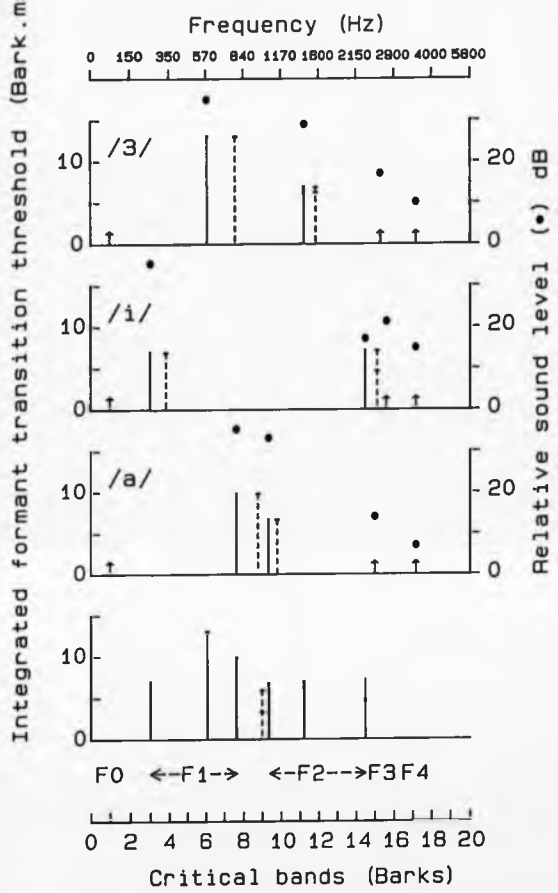

Frequency (Hz)
0 150 350 570 840 1170 1600 2150 2900 4000 5800
Integrated formant transition threshold
400
200
0
Rel. freq. diff. x time (%.ms)
(a)
4
2
0
Frequency diff. x time (kHz.ms)
(b)
20
10
0
Critical bands x time (Bark.ms)
(c)
←-F1-→
←--F2--→
0 2 4 6 8 10 12 14 16 18 20
Critical bands (Barks)

are plotted in more detail in Fig. 4.36 for the three vowel phonemes /a/, /3/ and /i/, and in a combined figure (lower). The solid vertical bars represent the integrated threshold values expressed in Bark.ms, obtained by averaging across duration of the relevant functions shown in Fig. 4.34. The three vowel sounds /3/, /a/ and /i/ are represented separately in the three upper diagrams and together, below. They are plotted as a function of critical band scale (Barks) to indicate spacings within which masking might occur. The dots and right-hand scales indicate the relative levels of the various formants. The positions of F0, F3 and F4 are indicated by the upward arrowheads. The dashed bars in the upper three diagrams indicate the highest upward shifts which occurred at the 10 ms duration. The lower horizontal bars on some lines represent the isolated-formant (IF) conditions. The dashed line on the lowest diagram represents the data of Gardner and Wilson (1979) normalised in the same way for downward and upward tone sweeps (the higher of the horizontal bars representing the threshold for downward sweeps).

From Fig. 4.35c, and its more detailed expansion in Fig. 4.36, there is now no overall trend with frequency, and the four-formant values all lie within a range of 2:1. The general implication of this is that when the threshold frequency increment is expressed in critical bands it is independent of frequency position and formant number. Thus the differences established above would appear to be due chiefly to frequency position. The mean value of this threshold across all four-formant stimuli is 8.44 critical band x ms. If one takes a 250 ms integrating time for the auditory system this indicates a possible asymptotic threshold of 0.034 critical bands, e.g. 5.4 Hz at 1 kHz.

This can be compared with the pure-tone frequency difference limen which ranges from 1 to 3 Hz at 1 kHz according to different

Fig. 4.36


Frequency (Hz)
0 150 350 570 840 1170 1600 2150 2800 4000 5800
/3/
/i/
/a/
10
0
20
Integrated formant transition threshold (Bark.ms)
Relative sound level (•) dB
F0 ←-F1-→ ←-F2--→F3F4
0 2 4 6 8 10 12 14 16 18 20
Critical bands (Barks)

authors and procedures. A more direct comparison, mentioned above, is with the results of Gardner and Wilson (1979). The mean values for upsweeps and downsweeps were 3.3 and 5.9 Bark.ms respectively. These are shown by the dashed line (Fig. 4.36, lower). It appears from this that changes within a speech-like sound are about twice as difficult to detect as within a pure-tone psychophysical context.

The integration time of the auditory system refers to its time constant. The time constant is the maximum time interval over which one event can interfere with a subsequent event (and is measured between the offset of the first and the onset of the second event). The time constant has been measured by Hughes (1946) as 250 ms and by Plomp (1964) as being between 200 and 300 ms. Sensory integration refers to a time band over which information from an acoustic waveform can be added together or combined in some way.

4.5.2 Arguments For Representing Thresholds on a Bark Scale

Where comparisons between four-formant and isolated, and upward and downward conditions were made, the representation of the thresholds was as unconverted "binfit" levels for the contrast analysis. These unconverted levels refer to the mean and standard deviation values produced by the analysis program. In Chapter 2.4 it was explained how the interval ratios between levels of difficulty were converted to values separated by a linear number scale for the purposes of computer analysis. The means and standard deviations produced by the program then needed to undergo a reverse transformation in order to be presented in meaningful, and for conditions of differing frequency, comparable units. All of the above conditions were carried out at the same reference frequency, and for reasons of greater accuracy (which are outlined in

Appendix B), contrast analysis could be carried out on the data without transforming it back into its original form.

Where the reference or start frequency was different, contrast analysis was employed on threshold values expressed as threshold frequency increment (i.e. Δf / f x 100 = %). The comparisons may be a fair reflection of differences in difficulty when the start or reference conditions are close to each other. If one looks at the frequency difference limen data of Moore (1973a) and (1974), and also Wier et al (1977) then one can see that frequency discrimination (with pure tone stimuli) is not a linear function over the speech range of frequencies, whether threshold is expressed in frequency extent (Hz) or percentage of start frequency. Assuming that frequency discrimination and frequency transition detection are closely related abilities, which has been the assumption made so far, then one would not expect frequency transition detection thresholds to be level as a function of frequency (when expressed in these terms) either. It is perhaps, therefore, worth repeating the contrast analyses where the start frequency differs between two stimulus conditions, but this time representing the thresholds in a way that is completely independent of frequency.

To a certain extent the expression of the thresholds in terms of percentage frequency change has a similar normalising effect as expressing them as critical bands. (This is because critical bandwidth increases with frequency). Fig. 4.35a demonstrates, however, that thresholds expressed in terms of percentage frequency change decrease systematically with increasing frequency. Therefore, contrast analysis performed on sets of thresholds expressed in this way will demonstrate a bias against conditions with a lower reference frequency.

4.6 Results 2

4.6.1 Contrast Analysis of Threshold Pairs Expressed in Terms of Critical Bands

As frequency transition detection thresholds appear to be independent of frequency when expressed in terms of critical bands, then it is appropriate to perform contrast analyses on pairs of thresholds which, due to different formant number or vowel identity, have different reference frequencies.

Figs 4.37 - 4.39 illustrate pooled transition detection thresholds for the first and second formants of the three test vowels. As before, for /3/ and /a/, the differences were significant ($p<0.001$, 1 or 2-tailed) that F2 transitions are easier to detect than those of the first formant.

However, for the vowel /i/, a contrast analysis of the thresholds for F2 and F1 transitions proved not to be significantly different. This differs markedly from the previous result where F1 transitions were found to be harder when expressed in terms of percentage frequency change.

If one compares Figs 4.37 - 4.39 with Figs 4.20 - 4.22 then in all cases there is a lessening of the difference between F1 and F2 transition detection thresholds when expressed in terms of critical bands.

The mean difference between the thresholds for F1 and F2 transitions decreases consistently the larger the frequency difference between the first and second formant. This may be a consequence of the particular three vowels used, but is more likely to be due to normalisation of the thresholds in terms of critical bands. In Fig. 4.35a one can see how systematically threshold

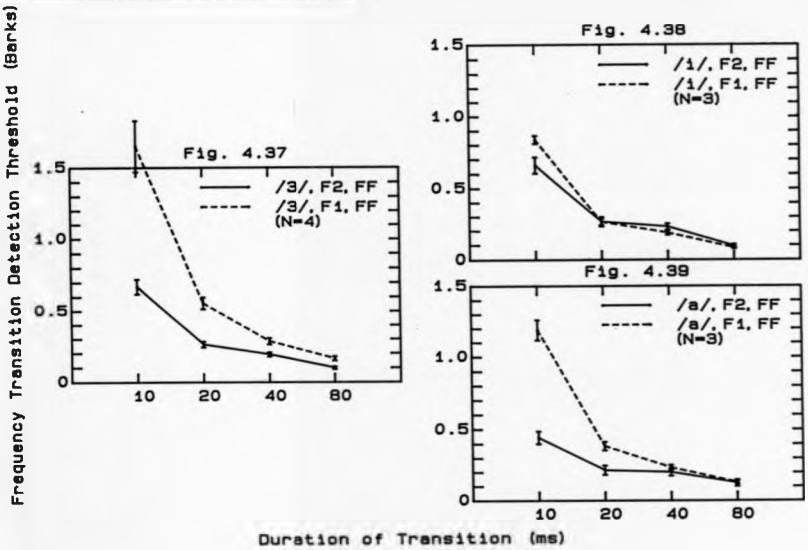


Fig. 4.37
/3/, F2, FF
/3/, F1, FF
(N=4)
Fig. 4.38
/i/, F2, FF
/i/, F1, FF
(N=3)
Fig. 4.39
/a/, F2, FF
/a/, F1, FF
(N=3)
Frequency Transition Detection Threshold (Barks)
Duration of Transition (ms)
0
0.5
1.0
1.5
10
20
40
80

decreases with increasing formant frequency when threshold is expressed in terms of percentage frequency change. One can see, conversely, in Fig. 4.35c how relatively even threshold becomes with frequency when expressed in terms of critical bands.

Fig. 4.40 illustrates a comparison between the rising F2's of /3/ and /i/. The result was significant, for one subject, ($p<0.01$, 1 or 2-tailed) that the rising F2 of /3/ is easier to detect than that of /i/. This is in the same direction as the difference between the two conditions when expressed as percentage frequency change, (see Fig. 4.23), but is more significant. From Figs 4.36 and 2.2 one can see how close the F3 of /i/ is to its F2 in frequency. Also, the F3 of /i/ has a larger amplitude; and at threshold for a 10 ms transition the F2 is within a critical band of the F3. Bearing in mind the even spacing of the formants of /3/ and for all the above reasons, the outcome of a contrast analysis between the two conditions would have been predicted.

Fig. 4.41 illustrates the detection thresholds of the F2 of /3/ and /a/. As was found previously (Fig. 4.24) there is no significant difference between the two conditions. The explanation of this result (see section 4.4) still holds as no difference was expected.

Figs 4.42 and 4.43 show a comparison between the F2 detection thresholds of /i/ and /a/ for two subjects. One subject (PC) found the F2 of /a/ easier to detect than that of /i/ ($p<0.01$, 1 or 2-tailed) as did the other subject (DAR) ($p<0.001$, 1 or 2-tailed). The results are the same as demonstrated previously in Figs 4.25 and 4.26, but for subject PC were more pronounced. The overall outcome was not predicted (either way) in the previous results section. However, there may be a difference due to the F2 of /a/ having a greater relative amplitude in relation to its surrounding formants than that of the F2 of /3/.

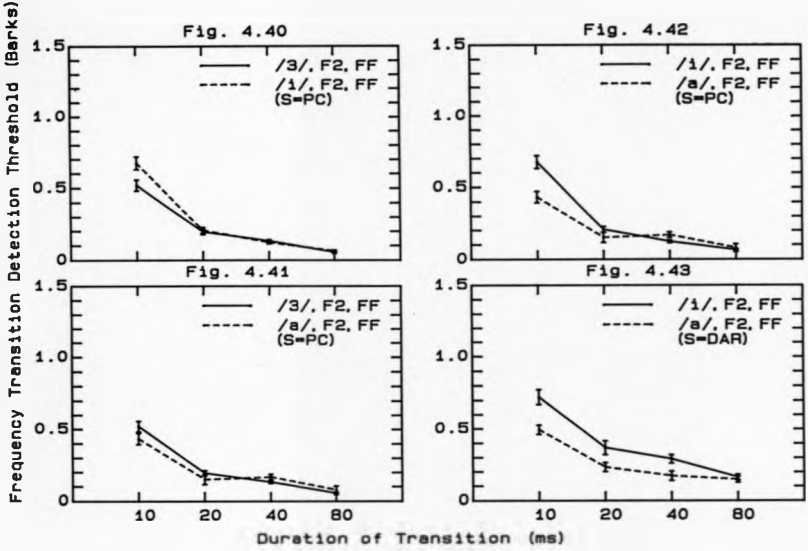


Fig. 4.40
/3/, F2, FF
/i/, F2, FF
(S=PC)
Fig. 4.42
/i/, F2, FF
/a/, F2, FF
(S=PC)
Fig. 4.41
/3/, F2, FF
/a/, F2, FF
(S=PC)
Fig. 4.43
/i/, F2, FF
/a/, F2, FF
(S=DAR)
Frequency Transition Detection Threshold (Barks)
1.5
1.0
0.5
0
10
20
40
80
Duration of Transition (ms)

Comparisons were also made between F1 conditions, expressing thresholds in critical bands. Fig. 4.44 shows the thresholds for the F1's of /3/ and /i/ for one subject. Previously, see Fig. 4.27, the F1 of /3/ was significantly easier than that of /i/; expressing the thresholds in critical bands caused a change in result in the opposite direction. This time it was significant ($p<0.001$, 1 or 2-tailed) that the F1 of /i/ is easier to detect than that of /3/. No particular explanation of this change in direction of result is offered apart from the fact that the thresholds for /3/ at 10 ms were abnormally high for two subjects (see Figs 4.10 and 4.11). This may well have adversely affected the result.

Fig. 4.45 shows a comparison between thresholds for the rising F1 of /3/ and that of /a/ for one subject. No significant difference was found between the two conditions.

Previously, as illustrated in Fig. 4.28, it was found that /a/ was easier to detect than /3/. If any prediction were to be made then one would perhaps expect the F1 of /3/ to be an easier condition. The proximity of the F2 of /a/ to its F1 is such that at 10 ms the frequency extent of the formant transition is less than a critical band (see Fig. 4.36). Nevertheless, expressing the thresholds in critical bands as opposed to percentage frequency change neutralises a previously unexplained result.

Finally, Figs 4.46 and 4.47 show individual comparisons between the F1's of /i/ and /a/. For one subject (PC) the result was significant ($p<0.001$, 1 or 2-tailed) that the F1 of /i/ was easier to detect than that of /a/. For the other subject (DAR) the result was also significant to a slightly lesser extent ($p<0.001$, 1-tailed; $p<0.01$ 2-tailed). This is in direct contrast to the previous results (Figs 4.29 and 4.30) which were significant in the opposite direction. For these two results then it seems likely that the F2 of

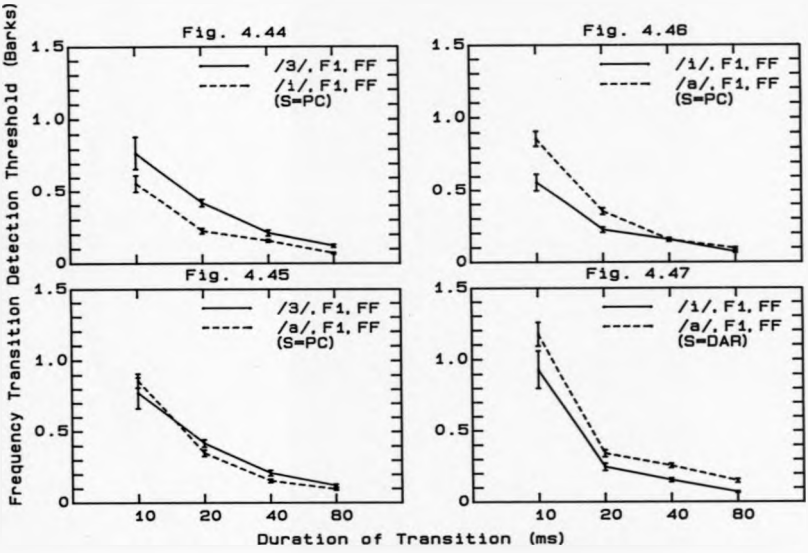


Fig. 4.44
/3/, F1, FF
/i/, F1, FF
(S=PC)
Fig. 4.46
/i/, F1, FF
/a/, F1, FF
(S=PC)
Fig. 4.45
/3/, F1, FF
/a/, F1, FF
(S=PC)
Fig. 4.47
/i/, F1, FF
/a/, F1, FF
(S=DAR)
Frequency Transition Detection Threshold (Barks)
1.5
1.0
0.5
0
10
20
40
80
Duration of Transition (ms)

/a/ has a downward masking effect on the F1 of /a/ as was suggested in the above paragraph.

4.6.2 Contrast Analysis of Threshold Pairs with the 10 ms Data Omitted

In Chapter 4.5.1 the concept of temporal integration and its implications for the present data set were discussed. Stephens (1973a) found that temporal integration is not optimal until upwards of about 15 ms. Abrupt changes in thresholds were observed in the frequency transition detection data between transition durations of 10 and 20 ms. Taken together these above two observations would suggest that it might well be sensible to repeat the contrast analyses between different stimulus conditions but this time omitting the data obtained with 10 ms transition durations. If, as seems likely, 10 ms is indeed too short for optimal temporal integration then inconsistent effects are possible over different formant, frequency and vowel conditions. In fact, due to the gating of stimulus offsets a good deal less than a 10 ms frequency transition is perceived in the shortest duration condition. The last 5 ms of each stimulus is multiplied by a Hanning window (as detailed in Chapter 2.1), and therefore the transition will cease to be perceivable before the end of the stimulus.

Contrast analysis was consequently performed on all of the above tested threshold pairs for transition durations of 20, 40 and 80 ms. In order to avoid repeating a lot of previous discussion the only results reported in the present section are where there is any change from the original analyses in sections 4.4 and 4.6.1.

In the pooled data comparisons there was only one change in result as a consequence of removing the 10 ms data. The difference

between the rising F2 of /3/ in isolated- and four-formant conditions now proved to be no longer significant. Originally, as illustrated in Fig. 4.18, the outcome was only marginally significant in favour of the isolated transitions being easier to detect ($p<0.05$, 1-tailed).

There were rather more changes in some of the individual data comparisons. The general trend, as indicated by the single pooled case above, was to reduce the value of 'z' and hence (sometimes) the level of significance between different vowel or formant conditions; i.e. the 10 ms data tended to emphasise (or even overemphasise) differences.

Briefly, for the upward transitions of the F2's of /3/ and /i/, the difference is no longer significant. This is not very surprising if one looks at the data in Fig. 4.40. The only marked difference between the two vowels occurred at the 10 ms transition duration. The differences between the F2's of /i/ and /a/ are now no longer significant, whereas originally (see Fig. 4.42) /a/ was the easier of the two to detect ($p<0.01$, 1 or 2-tailed).

The only result to become significant after not previously being so was for the comparison between the F1's of /3/ and /a/ (Fig. 4.45). It is easy to see why it is now significant ($p<0.001$, 1 or 2-tailed) that /a/ is easier than /3/ to detect. The 10 ms data was the only threshold pairing to show /3/ to be the easier of the two.

Finally, and possibly the most anomalous result in the light of current predictions, the comparison between the rising and falling F2 transitions of the vowel /i/ reveals an increase in significance. The original result, as illustrated in Fig. 4.32, was significant that rising transitions were easier to detect ($p<0.05$, 1-tailed). Removing the 10 ms data, however, made this much more emphatic

($p<0.001$, 1 or 2-tailed) and consequently more difficult to explain.

It was argued previously that the F3 of /i/ would have a downward masking effect on the F2 due to the nearness of the two formants. Therefore, one would expect rising transitions to be more difficult to detect. The incidence of anomalous results such as this one, with regard to predictions based upon influences of masking, requires that a more consistent method of representing (and analysing) the data be employed in order to justify a masking explanation of threshold differences between different conditions. If anything, removing the 10 ms data makes the general picture even less clear than before. The following section represents just such an attempt.

4.7 An Excitation Pattern Model of the Detection of Formant Frequency Transitions

4.7.1 Exposition of the Model

The original excitation pattern model was proposed by Zwicker (1956) and (1970). The current implementation, however, derives from Moore and Glasberg (1986b). Essentially, it predicts that a frequency change is detectable if the excitation patterns of two stimuli are different at any point by 1 dB (which is approximately the intensity difference limen for wideband noise, as outlined in Chapter 1.5). It must be pointed out that both Zwicker's original model, and Moore and Glasberg's implementation of it, were to explain the mechanism of frequency discrimination. The current implementation is designed to extend this by including data from frequency transition detection experiments.

To produce an excitation pattern of a waveform one must first

perform an auditory frequency analysis. A Fourier analysis, as illustrated for all three vowels in Fig. 2.2, typically shows narrow well-defined harmonic peaks. Although, it has to be noted that in the case of the figure just mentioned, the sampling rate and fundamental frequency were not integer powers of two; so a 512 point discrete Fourier transform's spectrum does not have points coinciding with the signal's harmonics. Therefore the peaks are not as narrow and well-defined as they would be with a signal designed with Fourier analysis in mind (e.g. with a fundamental of 128 Hz).

An auditory representation of the experimental stimuli is what is required in the present case. A Fourier analysis takes no account of the frequency selectivity of the auditory system as can be seen in Fig. 2.2 where the harmonics from 100 Hz to 5 kHz are all equally well resolved. An auditory representation, however, (as can be seen by the excitation patterns in Figs 4.50 - 4.52) shows that only the first few harmonics are resolvable. After that the auditory filter bandwidths become too wide to resolve the individual harmonic components, and increasing numbers of them are processed by individual channels as centre frequency rises.

The filterbank used for this analysis is discussed in detail in Patterson et al (1987). Briefly, the filter shape is based on a 'gammatone' function which is similar in many respects to the more established shape defined by Patterson (1976) (see Chapter 1.2.5). The advantage of the gammatone, apart from the fact that it is a simpler function computationally, (and therefore quicker to process on a computer) is that it provides a better approximation to the human auditory filter around the skirts than does the rounded exponential function of Patterson (1976).

The stimuli for input to the auditory filterbank were constructed in the same way as reported in Chapter 2.1. The

parameters for the transition stimuli were those of the obtained thresholds; i.e. the threshold value (in Hz) for a particular vowel and formant transition at a certain duration was used to set the frequency excursion for that condition. In the present case only the pooled data (as first shown in Figs 4.18 - 4.22) were analysed.

The hypothesis to be tested is that if masking fully accounts for differences between sets of data, then the difference between reference and test excitation patterns (with "threshold" stimuli as the test case) should show similar changes in intensity in relevant frequency channels.

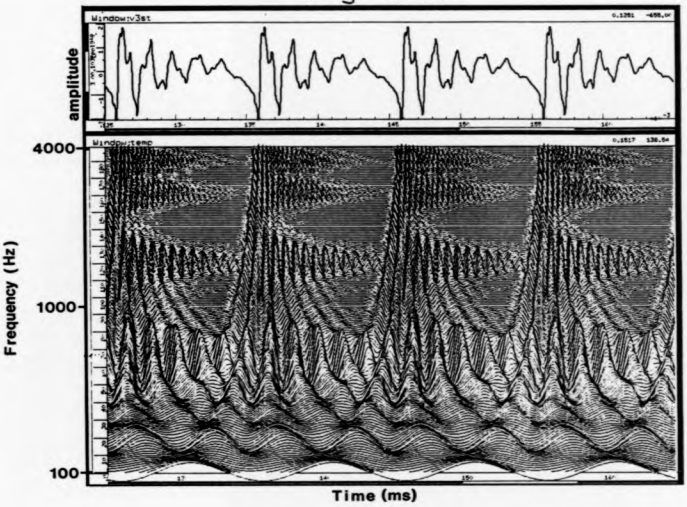
Fig. 4.48 shows the result of putting the steady-state vowel /3/ through an auditory filterbank. More channels are illustrated in this figure than are used in the analysis to give a better visual representation of the stimulus (and only four cycles, i.e. 40 ms, are shown for the same reason). Twenty-four channels were used to filter the synthetic vowels. This is the minimum number that will cover the frequency range (of between 100 Hz and 4 kHz) on an ERB scale. This passband was selected as it completely covered the range of the four formants of each of the vowels used. The centre frequencies of the filter channels were determined by the following two principles. Firstly, there is always a filter centered at 1 kHz. Secondly, the spacing between adjacent channels was calculated by the following function:

$$ERB = 6.23F^2 + 93.39F + 28.52$$

where F is frequency in kHz, and ERB is the equivalent rectangular bandwidth. The derivation of this equation comes from Moore and Glasberg (1983) who, in this paper, set out their method for calculating and producing excitation patterns.

It was then possible to calculate the power in each of the channels over a certain integration time, which was initially chosen

Fig. 4.48



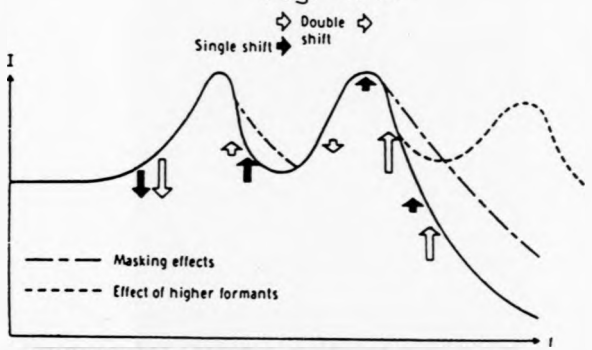
to be the time course of each formant transition. The resulting values for the reference stimuli were subtracted channel by channel from the corresponding threshold stimuli and then averaged over the three transition durations analysed. This was done in an attempt to further reduce any inherent variability in the data.

In the following section a 10 ms integration time is also used for the above method. A matrix of difference values are then observable over the time course of the formant transition. In comparing a steady-state 10 ms segment with a transition segment, which is progressively moved (in 5 ms steps) towards the end of the transition, it is possible to observe more discrete changes than in the initial method. By so doing it can be seen when criterion is achieved for each threshold stimulus. 10 ms was recommended by B.C.J. Moore (personal communication) as a suitable time period for such an analysis, more especially as it is equal to the duration of the glottal cycle, therefore not being out of phase with the periodic decaying of energy in the higher formants. The complete set of difference values obtained from both of the above methods are tabulated in Appendix C.

Fig. 4.49 is taken from Haggard (1977), and illustrates the consequences for the spectrum envelope of single and double formant changes. With a single upward formant transition, as for the stimuli in the current analysis, one sees a decrease in the low frequency slope of the moving formant; a rise is also observable on the high frequency side. Sometimes, as will be seen in the following results section, there is also an increase in energy in the neighbouring higher frequency formants. (Although this energy increase is typical of a serial synthesis system, as Haggard's figure illustrates, the same tendency is observable in some of the patterns shown below).

The 10 ms transition duration thresholds were omitted for the

Fig. 4.49



SPECTRUM ENVELOPE CONSEQUENCES OF FORMANT FREQUENCY CHANGES

reasons outlined in section 4.6.1 above.

4.7.2 Testing the Model

In the upper parts (a in each case) of Figs 4.50 - 4.52 the excitation patterns of the three vowels are illustrated. They can be regarded as an auditory representation of the Fourier transforms of Fig. 2.2. The ordinates in both cases are the same (relative level in dB), but the abscissae differ. The Fourier transform uses a linear number scale whereas the excitation patterns below are on an 'ERB' scale, which is quasi-logarithmic.

Below the excitation patterns are the 'difference' patterns constructed from temporally averaged threshold data, as detailed in section 4.7.1 above. These demonstrate, as did Haggard's figure (Fig. 4.49) the consequences of single formant movements on the power spectrum of a vowel. The formant conditions are varied enough, however, to show that Haggard's representation is only a specific case from a larger set of envelope consequences of single formant movement.

The difference patterns shown in Figs 4.50 - 4.52 correspond directly to the numbers in the right-hand columns of the first eight tables of Appendix C. As explained earlier, these values derive from rather arbitrary integration times, insofar as the reason for each time selected was that it was equal to the duration of the formant transition under investigation. This method of analysis, however, was used as a starting point for understanding the trends in the data, and how different vowel and formant conditions could be mapped onto each other. No attempt is made to justify this initial method in terms of psychological reality.

Starting with the patterns produced by F1 and F2 (in four- and

Fig. 4.50

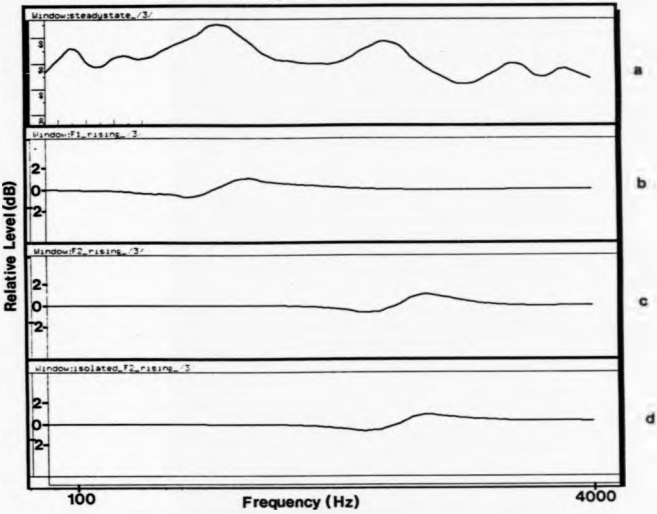

Window:steadystate_/3/
Window:F1_rising_ 3
Window:F2_rising_/3/
Window:isolated_F2_rising_ 3
Relative Level (dB)
100
Frequency (Hz)
4000
a
b
c
d

Fig. 4.51

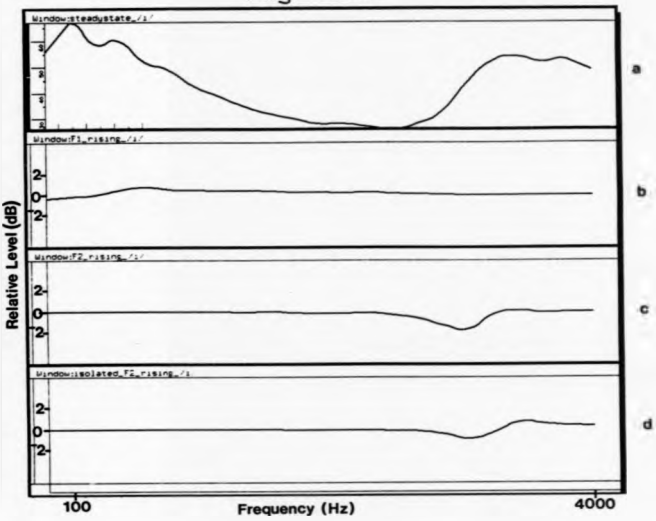
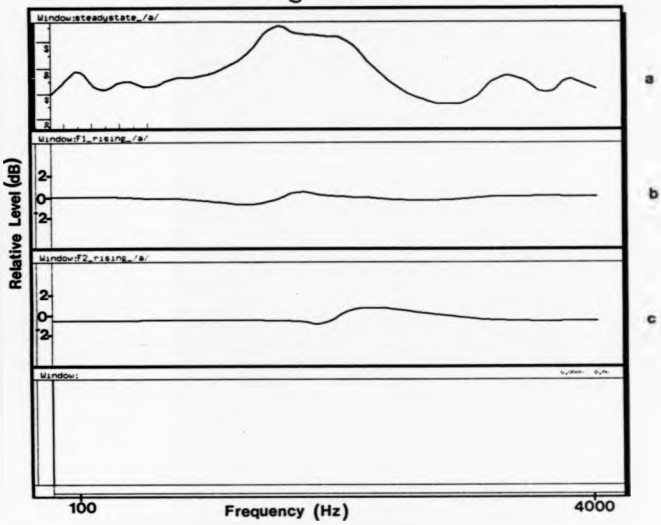

Window:steadystate_/s/
Window:F1_rising_/s/
Window:F2_rising_/s/
Window:isolated_F2_rising_/s
a
b
c
d
2
0
2
Relative Level (dB)
100
Frequency (Hz)
4000

Fig. 4.52


Window:steadystate_/a/
Window:F1_rising_/a/
Window:F2_rising_/a/
Window:
Relative Level (dB)
2
0
2
a
b
c
100
Frequency (Hz)
4000

isolated-formant context) in Fig. 4.50(b - d), we can see in each case the predicted pattern from Fig. 4.49. That is, a decrease in the low frequency side followed by a consequent level rise on the high frequency side. For the vowel /i/ (see Fig. 4.51) the picture is slightly different. The F1 transition produces the expected curve, but the F2 transition in four-formant context (c) just produces a negative peak without any corresponding positive shift on the high frequency side. The isolated F2 transition (d) does however contain the positive shift; this must be due, therefore, to the effect of the F3 of /i/ (in the four-formant condition). Its close proximity and comparable level (to the F2) possibly make any level changes small and dissipated.

Fig. 4.52 shows two difference patterns (as an isolated condition was not carried out) for the vowel /a/. The pattern for the F1 transition (b) is largely the same as in the majority of cases, i.e. alternately negative then positive moving from the low to the high frequency side of the first formant. The pattern then undulates in the region of the higher formants, but not as much as around F1 (and not in any channel above criterion levels). The final pattern, Fig. 4.52c, is the least characteristic of all the difference patterns so far produced. The majority of the twenty-four frequency channels are above criterion in this case, the rising F2 of /a/. The only area actually below criterion in this pattern is on the high frequency side of the second formant, but curiously, in the region of the low frequency slope of the third formant. No explanation can at present be given for this apart from, perhaps, that the F1 is far ranging in its influence on neighbouring frequency regions. As can be seen from the excitation pattern in Fig. 4.52a, the F1 of /a/ is both close to, and more intense than, the F2. Moving the second formant upwards, and therefore away from, the first

formant seems to have affected most of the difference pattern above and below the original F2 centre frequency. Possibly then the two resonances, being so close together, were interacting to an extent that they were raising the level over most of the signal's frequency range. The excitation pattern shows the two formants to be almost merged, so it is not unrealistic to propose that a (virtually single) broad-band high-level resonance could have such an effect as has been shown.

One way of testing the hypothesis outlined in section 4.7.1 (that the difference patterns of all the transition conditions should be similar if masking effects can adequately account for the data) is to integrate each pattern across frequency channels. It seemed reasonable to assume that the area of the difference patterns should be equal if the method is able to remove dependencies apparent in the original data. This assumption, however, proved not to be the case, the area calculations varying far more widely than any previous attempt at threshold normalisation.

However, the original aim of evaluating the data in terms of Zwicker's model proved, at least with the temporally averaged data, to be satisfactory. The right-hand columns of Tables 1 - 8 in Appendix C have the above criterion channels for each condition labelled with an asterisk. In the majority of cases there are three channels in the 1 dB range, but in order for one channel to reach criterion it is necessary to have corresponding changes in level in adjacent channels also. In only one channel out of all of the averaged data (Table 5 channel 19) is there a difference beyond 1 dB, and even then by only a marginal amount.

Tables 4.1 - 4.3, see below, illustrate the second method of viewing the difference pattern data as previously outlined in section 4.7.1. For each duration of transition successive 10 ms time slices

are integrated with 5 ms overlaps and differenced channel by channel with corresponding sections from a 'steady-state' excitation pattern.

Table 4.1 Level Difference Matrix for rising F2 of /3/: 20 ms transition

		time (ms) from start of transition		
channel	frequency	0	5	10
1	111	0.00	0.00	0.00
2	152	0.00	0.01	0.01
3	196	0.00	0.00	0.00
4	246	0.00	0.00	0.01
5	300	0.00	0.00	0.01
6	360	0.00	0.01	0.01
7	427	0.00	0.01	0.01
8	500	0.00	0.00	0.01
9	581	0.00	0.00	0.00
10	670	0.00	0.00	0.00
11	769	-0.01	-0.02	-0.03
12	878	-0.04	-0.06	-0.11
13	1000	-0.10	-0.14	-0.30
14	1135	-0.31	-0.35	-0.61
15	1286	-0.20	-0.92	-1.30
16	1454	0.08	0.23	-0.05
17	1642	0.70	1.02	1.75
18	1852	0.56	0.61	1.57
19	2089	0.26	0.27	0.78
20	2356	0.06	0.06	0.19
21	2657	0.00	0.00	0.00
22	2999	-0.04	-0.04	-0.09
23	3389	-0.01	-0.02	-0.03
24	3836	0.00	0.00	0.02

Table 4.2 Level Difference Matrix for rising F2 of /3/:

40 ms transition

time (ms) from start of transition

channel	frequency	0	5	10	15	20	25	30
1	111	0.00	0.00	0.00	0.00	0.01	0.01	0.00
2	152	0.00	0.00	0.00	0.01	0.01	0.00	0.01
3	196	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	246	0.00	0.00	0.00	0.00	0.00	0.00	0.01
5	300	0.00	0.00	0.00	0.00	0.01	0.01	0.01
6	360	0.00	0.00	0.01	0.01	0.01	0.01	0.01
7	427	0.00	0.00	0.00	0.01	0.01	0.01	0.01
8	500	0.00	0.00	0.00	0.00	0.00	0.01	0.01
9	581	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	670	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01
11	769	0.00	-0.01	-0.01	-0.01	-0.02	-0.04	-0.04
12	878	-0.01	-0.02	-0.04	-0.06	-0.09	-0.10	-0.12
13	1000	-0.04	-0.05	-0.11	-0.13	-0.20	-0.22	-0.29
14	1135	-0.12	-0.14	-0.31	-0.34	-0.52	-0.54	-0.55
15	1286	-0.07	-0.34	-0.46	-0.83	-1.01	-1.35	-1.60
16	1454	0.03	0.11	-0.03	-0.07	-0.19	-0.26	-0.31
17	1642	0.24	0.35	0.69	0.82	1.22	1.37	1.67
18	1852	0.19	0.20	0.52	0.55	0.90	0.92	1.31
19	2089	0.09	0.10	0.25	0.26	0.43	0.44	0.65
20	2356	0.02	0.02	0.06	0.06	0.09	0.09	0.15
21	2657	0.00	0.01	0.00	0.00	0.00	0.00	0.00
22	2999	0.00	0.00	-0.03	-0.03	-0.06	-0.06	-0.08
23	3389	0.00	0.00	-0.01	-0.01	-0.03	-0.03	-0.03
24	3836	-0.01	0.00	0.01	0.01	0.00	0.00	0.03

Table 4.3 Level Difference Matrix for rising F2 of /3/:

80 ms transition

time (ms) from start of transition

ch	freq	15	20	25	30	35	40	45	50	55	60	65	70
1	111	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00
2	152	.00	.00	.00	.00	.00	.00	.01	.01	.00	.00	.01	.01
3	196	.00	.00	.00	.00	.00	.00	.00	.01	.01	.01	.00	.00
4	246	.00	.01	.00	.00	.00	.00	.01	.01	.00	.00	.00	.01
5	300	.00	.00	.00	.00	.01	.00	.01	.01	.01	.01	.01	.01
6	360	.00	.01	.01	.00	.01	.01	.01	.01	.01	.02	.02	.02
7	427	.00	.00	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01
8	500	.00	.00	.00	.00	.01	.01	.01	.01	.01	.01	.01	.01
9	581	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
10	670	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	-.01	-.01
11	769	.00	-.01	-.01	-.02	-.02	-.02	-.02	-.03	-.03	-.03	-.03	-.03
12	878	-.02	-.03	-.05	-.05	-.06	-.07	-.07	-.08	-.08	-.08	-.08	-.08
13	1000	-.06	-.08	-.09	-.11	-.13	-.16	-.17	-.18	-.18	-.18	-.19	-.19
14	1135	-.15	-.22	-.23	-.30	-.32	-.40	-.41	-.45	-.45	-.45	-.45	-.34
15	1286	-.33	-.38	-.54	-.61	-.77	-.84	-.98	-1.03	-1.08	-1.08	-1.09	-1.17
16	1454	.00	-.07	-.10	-.16	-.18	-.22	-.24	-.29	-.30	-.31	-.32	-.34
17	1642	.33	.47	.52	.68	.74	.90	.96	1.05	1.07	1.06	1.06	.99
18	1852	.22	.34	.34	.49	.50	.66	.66	.74	.75	.74	.74	.75
19	2089	.13	.17	.17	.26	.26	.34	.34	.35	.35	.36	.36	.36
20	2356	.02	.03	.03	.06	.06	.07	.07	.08	.08	.07	.07	.08
21	2657	.00	.00	.00	.00	.00	.01	.01	.00	.00	-.01	-.01	-.01
22	2999	-.01	-.03	-.03	-.03	-.03	-.04	-.04	-.04	-.04	-.05	-.05	-.06
23	3389	-.01	-.02	-.02	-.01	-.01	-.02	-.02	-.01	-.01	-.01	-.02	-.03
24	3836	-.01	.01	.01	.00	.00	.00	.00	.02	.02	.01	.01	.01

The above three tables show only the one condition (out of the eight considered). The rest of the tables are in Appendix C (Tables 9 - 29). This is too great a number to present in the current discussion, so a representative case (in the rising F2 of /3/) has been selected for consideration.

In the 80 ms transition condition, both here in Table 4.3 and in Appendix C, only durations after the first 15 ms are shown. This is for two reasons. Firstly, the majority of the columns for the first three transition durations were populated by zeros, and secondly, it is not possible to display the information for one channel in a single row otherwise (even if all superfluous spacings are removed as is done here).

The underlining of a value denotes an above criterion channel. As one would expect, in the frequency regions of interest the difference values steadily grow until criterion is achieved. However, as can be seen in all the 80 ms cases, criterion is always achieved some way before the end of the formant transition. The shorter durations, as represented in Tables 4.1 and 4.2, go some way beyond criterion in the relevant frequency channels. In the previous method averaging over the three transition durations ensured that this was not the case. However, as mentioned in section 4.5.1, the gating of the last 5 ms of the stimulus would have detrimental effects on the perception of a frequency transition. This would be of even greater importance for the shorter duration transitions.

The above method is useful in enabling the observation of discrete changes in excitation patterns and provides some insight into how the auditory system may detect changes in frequency. In Chapter 1.4.6 both Horst (1982) and Lacerda and Moreira (1982) were quoted as saying that the auditory system treats frequency transitions as sequences of short stationary segments. Therefore,

accepting this precept, it seems possible that the auditory system could (using a constant integration time such as 10 ms) do the sort of sampling as illustrated in the above three tables.

4.8 Discussion 2 and Conclusion

The results reported here indicate that frequency transition detection threshold is inversely proportional to transition duration over 10 - 80 ms and therefore lends no support for the existence of specific rate-of-change detectors (for which the opposite trend would have been expected).

Threshold is approximately constant over frequency when expressed in critical bandwidths (Barks) at that frequency, and therefore lends no support for the existence of specific formant detectors. In percentage frequency terms, however, F2 does appear to have a lower threshold than F1. This could be due, in part, to FTC's at lower frequencies being comparatively broader and more symmetrical (Evans, 1972). The implication of this is that sounds are more easily masked at lower frequencies. At higher frequencies, where FTC's are narrower with sharp cut-offs on the high frequency side, maximal masking effects are produced with maskers of a lower frequency than that of the maskee.

Perhaps surprisingly, the detection of a transition within one formant of a four-formant complex is only about twice as difficult as the tasks of frequency discrimination and the detection of a pure-tone frequency sweep. This will be tested in a more direct comparison in the next chapter.

It may be argued that taking an average of such wide-ranging, and in some cases seemingly conflicting data, (see Fig. 4.34) is not particularly advisable. In order to establish whether any general

trends were present in the data, however, such methods were necessary.

The effect of masking was the reason for using the vowels /i/ and /a/ in the present study. /i/ was chosen for downward masking effects (of F3 upon F2) and /a/ for similar reasons. It was always assumed that the proximity of a neighbouring formant hinders rather than helps the perception of a formant transition. On the whole this is borne out by the results of contrast analysis on different vowels in similar formant conditions.

In order to further establish if explaining the results in terms of masking effects Zwicker's excitation pattern model was employed. Although integrating difference patterns across frequency channels proved to be ineffective (in terms of normalising the data), the evaluation of the patterns with respect to whether criterion had been achieved proved satisfactory. The former method was always going to be difficult with so few data points providing little leeway in the threshold values. It must not be forgotten that the threshold data came (in the case of the three vowel) from different subject groups, with presumably different abilities.

The modelling of the data, purely in terms of the 1 dB criterion, appears to support the view that differences in threshold between vowel and formant conditions can be accounted for in terms of equal energy changes in excitation patterns. The successive integrations of short time slices of stimulus revealed how the auditory system may process the stimulus in order to decide if a frequency transition has in fact taken place.

CHAPTER 5

Frequency Transition Detection Using Pure Tones

5.1 Introduction

In the previous two chapters tentative comparisons have been made between results for formant frequency transition detection thresholds in isolated vowels and thresholds for the same task obtained by other workers using pure-tone stimuli. The reason that pure-tone data are used for comparison with these studies is that the sine wave, as the simplest possible acoustic signal, provides one with the lowest thresholds for such a task. This then allows comparison with data from stimuli of more complexity and enables one to look at the effect of greater spectral content on frequency transition detection. Pure tones are also a case where simultaneous masking, for a frequency transition detection task, does not exist; even with well-spaced formants, as in the vowel /3/, masking cannot be completely discounted.

Pure-tone frequency discrimination data have also been used for comparison. This is due to more data being available for pure-tone frequency discrimination; data for frequency transition detection of pure tones have not been as systematic or detailed. In view of the difficulty of such comparisons, due to variations in stimulus frequencies and durations, a set of experiments were carried out in similar conditions to those of Chapter 4.

This enabled direct and exact comparisons to be made between formant frequency transition and sinewave transition detection thresholds. An additional bonus of this approach is that the methods employed in all of the frequency transition detection experiments can

be assessed for reliability and comparability. Frequency discrimination is a well documented area of psychoacoustics and is assumed to be a similar ability to that of frequency transition detection. Comparisons are made between the data obtained in the present chapter and two well-known studies of pure-tone frequency discrimination. The author found no appropriate studies in the frequency transition literature for comparison with these results. Comparisons are made between four-formant, isolated-formant and sinewave conditions in an attempt to ascertain whether differences in threshold can be explained purely in terms of degree of complexity of the stimulus.

5.2 Methods

As in the previous chapter, all stimulus durations were 200 ms, and transition durations of 10, 20, 40 and 80 ms were used within the test stimulus. The transitions were always in the final part of the test stimulus as before.

It was suggested in the previous chapter that, due to sensory integration, it was perhaps not necessary to use all of these durations in subsequent experiments. Unfortunately, it is necessary to use all four transition durations in order that appropriate comparisons can be made between the two sets of data. In section 5.5 additional reasons for the use of more than one transition duration are presented. Three test frequencies were chosen for the following experiments. These were 250 Hz (F1 of /i/), 1420 Hz (F2 of /3/) and 2320 Hz (F2 of /i/). The above three were selected to cover the range of formant frequencies tested in Chapter 4. The lowest and highest of these frequencies were also the extremes of formant frequency used for frequency transition detection in the previous

chapter. 1420 Hz is the test frequency closest to half-way between these two extremes.

Due to a (predictably) greater sensitivity to pure tone than formant transitions slower rates of frequency change were used in the present study. Rates of between 0.0225 and 11.3 Hz/ms were employed to cover the above three frequencies and the four fixed transition durations.

The stimuli were presented at an overall sound level of 75 dB SPL. This was the average sound level of presentation for the first and second formants of the three vowels used in the previous chapter.

The main purpose of the present chapter, as mentioned in the introduction, was to produce frequency transition detection data of pure-tone stimuli using an identical design to that of Chapter 4.

5.3 Results

Frequency transition detection thresholds are illustrated in Figs 5.1, 5.3 and 5.5 for pure tones and their respective isolated vowel (and formant) counterparts. The thresholds are expressed in terms of critical bands to be consistent with the representation in the previous chapter. The figures listed above all show a decreasing threshold with increasing transition duration. This is consistent with the thresholds for isolated vowels and formants, and also other workers (Watson and Gengel, 1969; Stephens, 1973b).

Figs 5.2, 5.4, and 5.6 show threshold ratios of the three possible pairings between four-formant, isolated-formant and sinewave conditions. Expressing the results in this way removes any dependence on duration and highlights any real differences between conditions. Fig. 5.2 illustrates these threshold combinations for

Key to Chapter 5 Figures

F1 = first formant

F2 = second formant

FF = four formant

IF = isolated formant

SW = sinewave

U = upward transition

D = downward transition

N = number of subjects

S = initials of subject

(for individual data)

FTD= frequency transition detection

Fig. 5.1

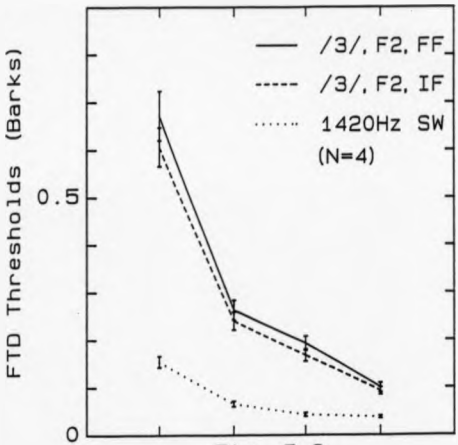

/3/, F2, FF
/3/, F2, IF
1420Hz SW
(N=4)
FTD Thresholds (Barks)
0.5
0

Fig. 5.2

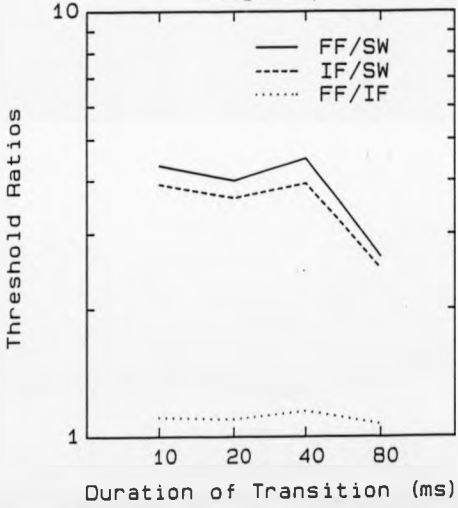

FF/SW
IF/SW
FF/IF
Threshold Ratios
10
1
10
20
40
80
Duration of Transition (ms)

Fig. 5.3

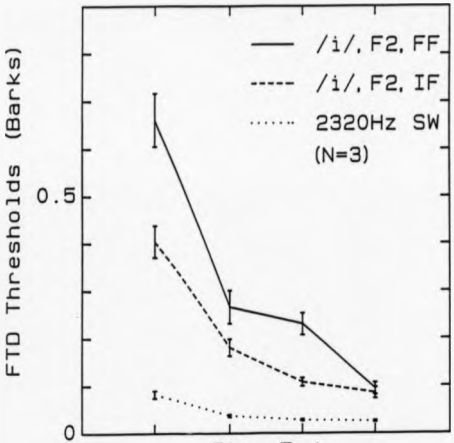

/i/, F2, FF
/i/, F2, IF
2320Hz SW
(N=3)
FTD Thresholds (Barks)
0.5
0

Fig. 5.4

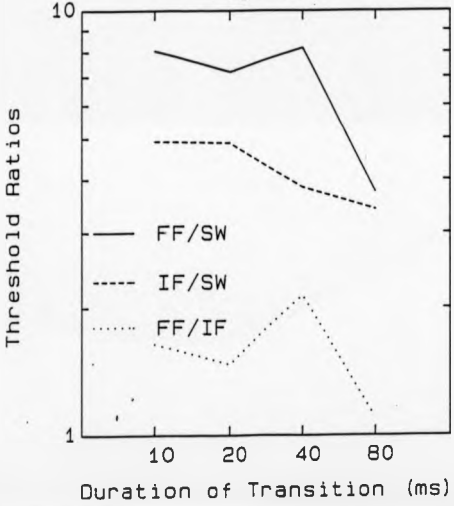

10
1
FF/SW
IF/SW
FF/IF
Threshold Ratios
10
20
40
80
Duration of Transition (ms)

Fig. 5.5

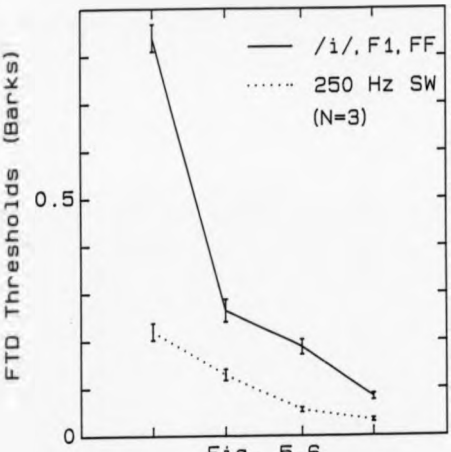
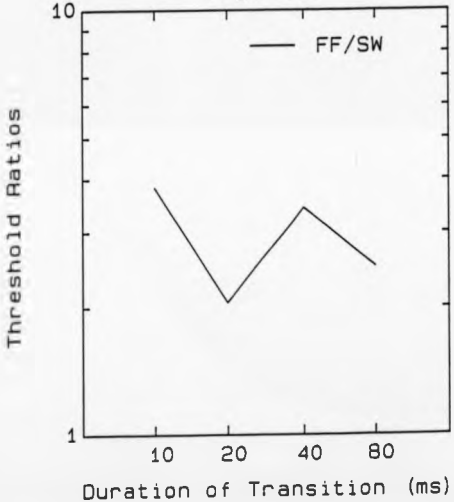

/i/, F1, FF
250 Hz SW
(N=3)
FTD Thresholds (Barks)
0.5
0

Fig. 5.6


FF/SW
Threshold Ratios
10
1
10
20
40
80
Duration of Transition (ms)

1420 Hz (i.e. the F2 of /3/). The functions produced by both the four-formant and isolated-formant vowel divided by the sinewave contrast sharply with the four-formant / isolated-formant ratio values. One may conclude from the resulting near-unity horizontal line that the perceptual criteria used for task of frequency transition detection for the second formant of the vowel /3/ is the same in both four- and isolated-formant conditions. In comparison with the sinewave data the functions for four- and isolated-formant conditions are much less regular.

Fig. 5.4 shows similar ratios to those of Fig. 5.2 but this time at 2320 Hz (i.e. the F2 of /i/). In this case all of the ratios describe differing functions. The four-formant / isolated-formant ratio function is not a horizontal line, and neither is it monotonic. The standard deviation values were left out of the ratio figures for the sake of clarity. When non-horizontal is referred to therefore, it should be taken that even including standard deviations, it is not possible to construct a line of zero gradient through the points. This could be taken as evidence of the masking effect of the F3 of /i/ upon a rising F2; one would have expected a horizontal straight line with a ratio of greater than one if there was no effect of neighbouring formants. The isolated-formant / sinewave ratio function is not horizontal but is monotonic. A descending function with increasing duration could be due to temporal integration starting to break down at the shortest transition duration (Stephens, 1973b), but having divided out the influence of duration by taking ratios it is difficult to support this view.

The reason for an undulating function for the four-formant / sinewave ratios in Figs 5.4 and 5.6 (for the F2 and F1 of /i/ frequencies respectively) is a little difficult to explain. It is a result of the gradient between the data points not becoming

consistently smaller with increasing duration for the four-formant stimuli. In both cases there is an increase in the gradient between 40 and 80 ms. For tone glides, in all cases, the gradient gradually decreases with increasing duration; this is consistent with Collins and Cullen (1978) who produced similarly contoured functions for tones over the 5 to 120 ms range of durations.

In Fig. 4.34 when formant transition thresholds were integrated with respect to transition duration, in certain cases similarly shaped functions resulted. Nabelek and Hirsh (1969), in some experiments to determine rate discrimination for tone bursts of changing frequency, found that for virtually all frequency intervals and frequencies between 250 and 4000 Hz that the best discriminability was for transition durations of around 30 ms. Whether glide rate discrimination and glide detection are directly related it is not clear, but it could partially account for the dips in the function between 20 and 40 ms for temporally integrated thresholds in Fig. 4.34 and also similar dips in Figs 5.2, 5.4 and 5.6 (where the ratio cancels out the effect of duration).

Table 5.1 summarises the ratio data by giving the values of mean ratios for each of the possible condition pairings.

Table 5.1 Mean Ratios

Frequency (Hz)	250	1420	2320
FF/SW	2.9	3.9	6.8
IF/SW	-	3.5	4.3
FF/IF	-	1.1	1.6

The table illustrates a tendency towards a more marked difference between conditions with increasing frequency in all ratio permutations. One should be careful not to overgeneralise such a

trend, no matter how consistent, from so few figures.

As with the previous chapter some subjects completed trials on more than one stimulus condition, and therefore made some individual comparisons between thresholds at different frequencies possible. The thresholds, as before, are normalised in terms of the critical bandwidth at each test frequency.

Fig. 5.7 illustrates pooled data thresholds for frequency transitions at both 250 and 2320 Hz. Contrast analysis showed clear significance ($p<0.001$, 1 or 2-tailed) that 2320 Hz sinewaves are easier to detect than those at 250 Hz. Individual tests for each of the three subjects from the above pooled condition were similarly significant (results not illustrated).

Fig. 5.8 shows a comparison for one subject (PC) between 1420 Hz and 250 Hz sinewaves. As with the previous figure, the higher of the two frequencies was significantly easier to detect ($p<0.001$, 1 or 2-tailed). The crossover of data points at 40 ms is slightly odd, and no explanation is offered for this trend. Fig. 5.9 shows the last of the individual comparisons between different reference frequency data (at the frequencies 1420 and 2320 Hz). As with the previous two results, the higher frequency is easier to detect ($p<0.001$, 1 or 2-tailed). So it seems that even with normalisation of the data in terms of critical bands that, over the range of 250 to 2320 Hz, the higher frequency is easier to detect for pure-tone stimuli. This is supported by Moore (1974) who published pure-tone frequency DL data that, when expressed as a function of critical bandwidth, was by no means constant.

Moore argued that the assertion by some workers (e.g. Zwicker, 1970; Scharf, 1970) that the frequency difference limen is a constant fraction of the critical bandwidth (1/27), was incorrect. Moore

Fig. 5.7

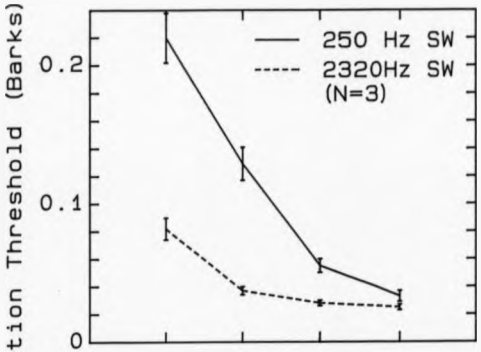

250 Hz SW
2320Hz SW
(N=3)
0.2
0.1
0

Fig. 5.8

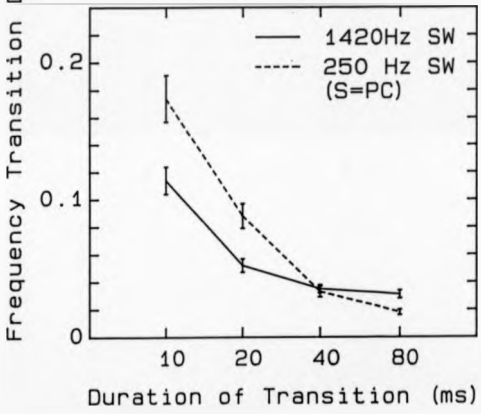

1420Hz SW
250 Hz SW
(S=PC)
0.2
0.1
0
10
20
40
80
Frequency Transition Detection Threshold (Barks)
Duration of Transition (ms)

Fig. 5.9

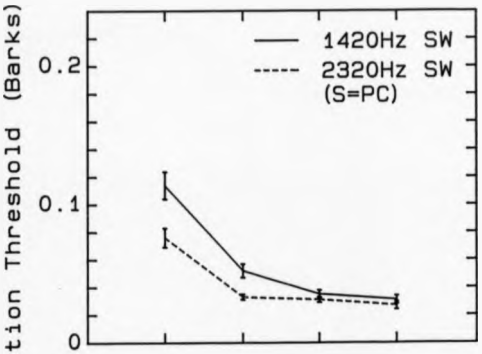
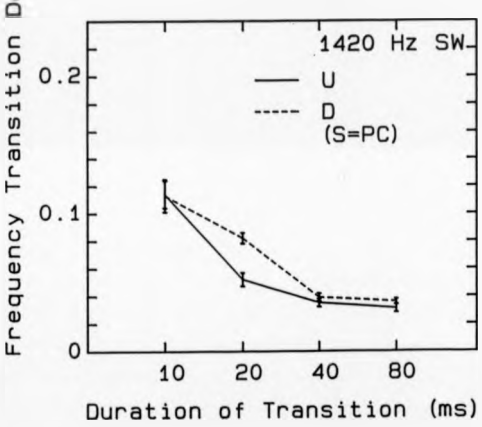


Fig. 5.10



showed that the assertion was based on data all of which was obtained by using modulation methods (e.g. Shower and Biddulph, 1931). Henning (1966) demonstrated the unreliability of the modulation method due to intensity fluctuations being available as an additional cue for discrimination. Moore (1974) published frequency DL data that, when expressed in terms of critical bandwidth (CB / ΔF) produced a descending set of values with increasing frequency. The present frequency transition data when expressed in similar terms is not consistent, but as demonstrated above is also not constant. The present results will be compared with other workers' data in the following section.

Finally, Fig. 5.10 illustrates upward and downward pure-tone frequency transition detection thresholds for one subject (PC). The differences between the two conditions did not prove to be significant but there is clearly a slightly greater sensitivity to upsweeps. This agrees with other workers results for tone bursts, notably Collins and Cullen (1978), Nabelek (1978) and Gardner and Wilson (1979).

The mean values for Gardner and Wilson's pure-tone up and downsweeps are 3.3 and 5.9 bark.ms (for a duration of 62.5 ms). If the present data are averaged over the four durations tested then the mean values are 1.51 and 1.67 bark.ms for upward and downward frequency sweeps respectively. If averaged over the closest durations to 62.5 ms (i.e. 40 and 80 ms) these figures rise to 1.93 and 2.24 bark.ms. They are still quite a lot lower than Gardner and Wilson's thresholds, though their data was taken from tones at 1000 Hz whereas the present data is for 1420 Hz. We have seen a tendency in the above data for thresholds to be lower at higher frequencies even when expressed in terms of critical bands, so this could account for some of the difference between the two sets of

data. It should also not be forgotten that the present data comes from only one subject and Gardner and Wilson's from just two, so it would be foolish to attach too much weight to the threshold differences.

5.4 Contrast Analysis of Threshold Pairs with the 10 ms Data Omitted

In the corresponding section of the previous chapter (4.6.2) only the cases where a change in significance between analyses of all four durations against the longest three were noted. This rationale is maintained here also.

Of all the contrast analyses reported in the previous section there is only one change in significance on removing the 10 ms data. This is in the comparison between 1420 and 250 Hz sinewaves for one subject (as illustrated in Fig. 5.8). Conforming with the general trend in the previous chapter, that is of reducing the difference between conditions when the 10 ms data have been omitted, the result moves from high ($p<0.001$, 1 or 2-tailed) to marginal significance ($p<0.05$, 1-tailed) that the higher frequency is easier to detect.

5.5 Comparison of Results

As mentioned in the introduction to this chapter, pure-tone frequency discrimination is a well-documented area of psychoacoustics. Two studies have been chosen for comparison with the current data, and they are Wier et al (1977) and Moore (1973a). The reasons for selecting the above two studies are as follows:-

1) They are both well-known and review other data.

2) The presentation of results was such that current data could be directly compared in both figure and table form.

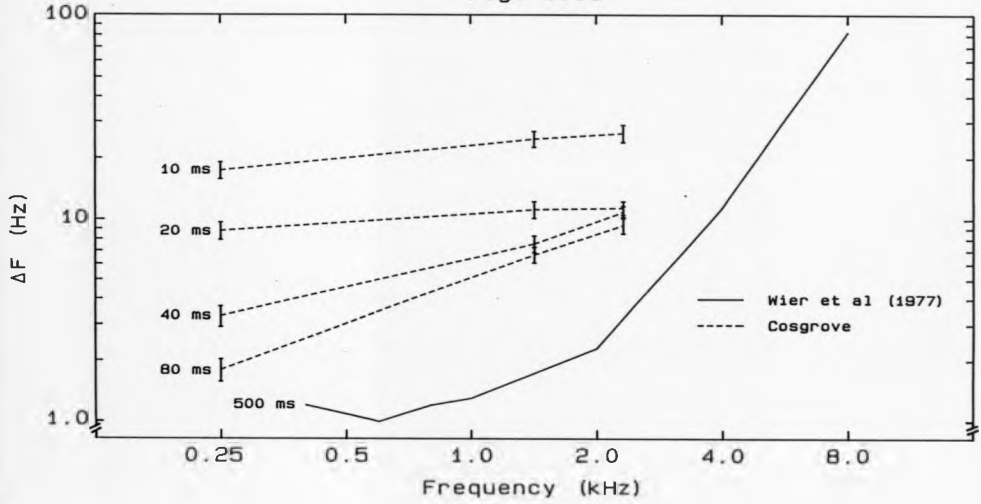
The data from the current study were from one subject (PC) who completed all of the three frequency conditions tested.

Fig. 5.11 shows some frequency discrimination data from Wier et al (1977). The stimuli for this study were 500 ms square-gated tone bursts presented at sensation levels between 5 and 80 dB SL, whereas the pure-tone frequency transition detection stimuli were presented at 75 dB SL. Wier et al's data are the mean of four subjects and the current data, as mentioned above, are from one subject only. The thresholds in Fig. 5.11 are expressed in terms of extent of frequency change (ΔF), and are quite apparently frequency dependent.

The differences between the two sets of data are largely due to stimulus duration, the longest duration of the current data being less than a sixth of Wier et al's tone-bursts. Wier's subjects were trained for at least 20 hours before experimental data were collected; the current study, in every chapter, employs only partly trained subjects. Finally, in the current study thresholds were obtained at the 75 % correct level of performance whereas Wier et al's thresholds were estimated at the 71 % level.

The difference in percentage correct level of performance would make the frequency DLs from Wier et al (1977) comparatively smaller than the frequency transition detection thresholds of the current study, as would the difference in amount of training given. The difference between total duration of stimuli arguably produces the greatest amount of difference between sets of thresholds. From Fig. 5.11 one may speculate that Wier et al's data would produce similarly contoured and positioned curves if the duration of the tone

Fig. 5.11


100
10
1.0
ΔF (Hz)
10 ms
20 ms
40 ms
80 ms
500 ms
Wier et al (1977)
Cosgrove
0.25
0.5
1.0
2.0
4.0
8.0
Frequency (kHz)

bursts were equal to one of those used in the current study.

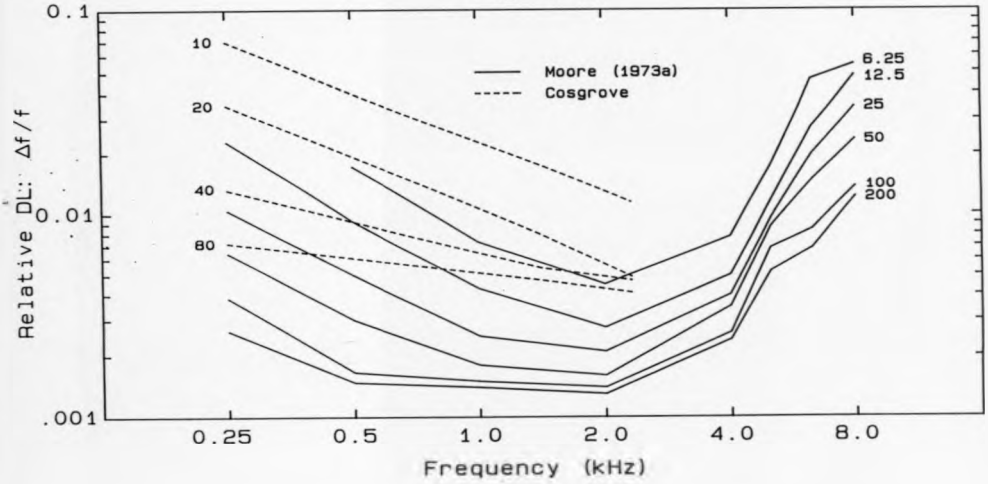
Fig. 5.12 illustrates pure-tone frequency DLs from Moore (1973a) and the same data from the current study as shown in the previous figure. The error bars are removed this time so as not to obscure detail. The data are expressed as Relative DL (i.e. Δf / f) and are plotted against test frequency.

The stimulus durations are far more comparable in this figure, encompassing all the durations used in the current study. Moore used a 200 ms interval within a stimulus pair whereas 500 ms was used in this study. Moore's stimuli were gated at onset and offset for a duration of 2 ms; 5 ms was chosen for the current stimuli. A 75 % correct level of performance was chosen for threshold estimation in both studies. Constant loudness was used by Moore for his stimuli whereas the current study employed a constant SPL. Moore presented his stimuli at 60 dB SPL at 1 kHz as opposed to 75 dB SPL for all frequencies in this study. Both sets of data are for one subject only. The subject chosen from Moore's study for illustration in Fig. 5.12 was the one (TC) who had completed trials in the majority of the experimental conditions.

Both sets of data produce similarly decreasing functions between 250 Hz and 2 kHz. On average Moore's data indicate a considerably lower threshold for comparable conditions. They also indicate that the present linear functions are unlikely to continue much above 2 kHz.

Equiloudness contours, as detailed in Robinson and Dadson (1956), show a minimum between 3 and 4 kHz. Even though Moore has attempted to eliminate loudness effects by presenting his stimuli at constant loudness, the current data are still decreasing at a reasonably sharp rate in relative threshold at 2320 Hz which suggests that the trend is a real one. In fact Moore's data, despite their

Fig. 5.12


0.1
0.01
.001
Relative DL: Δf/f
10
20
40
80
Moore (1973a)
Cosgrove
6.25
12.5
25
50
100
200
0.25
0.5
1.0
2.0
4.0
8.0
Frequency (kHz)

loudness normalisation, follow the same shaped curve as equiloudness contours over the frequency region tested. The current data are not normalised with respect to loudness which may explain the greater comparative gradient at 2 kHz.

From Fig. 5.12 we can see that the data from frequency DL and frequency transition detection experiments are reasonably comparable to each other. Obviously, from this and previous studies the picture is one of frequency discrimination being the more sensitive of the two abilities.

In order to try and make the current data 'sit' on top of that of Wier et al (1977) and Moore (1973a), or at least follow it a little more closely, a number of transformations were performed on the above thresholds. Firstly, all the data were integrated with respect to time. Sensory integration appears to occur in the vowel data of the previous chapter and, according to Hughes (1946), for all signals below 250 ms duration. Temporal integration produced ascending thresholds with respect to frequency in all cases (bar some of the lower frequencies and shorter durations in Moore's data) when the data was expressed in Hz.ms ($\Delta f \times d$). Generally, thresholds were larger for shorter durations.

If the data were temporally integrated for a second time (i.e. $\Delta f \times d^2$), and expressed as $Hz.ms^2$, then most of the inconsistencies in trend in Moore's and the current data disappears. The thresholds generally increase with increasing frequency and likewise with decreasing duration.

As temporal integration alone failed to normalise any of the data, or bring the current data any closer to that of the other two studies, other means of normalisation were sought. In fact, the data from Wier et al (1977) became more distant from the other two sets of

data with the temporal integrations, presumably due to the tone burst durations of the study being beyond the time constant of the auditory system (see Chapter 4).

The next attempt at data 'normalisation' was that of expressing the thresholds in terms of critical bands (i.e. Δf / CB). This had the effect of producing reasonably level functions in the current data at the 40 ms duration of frequency transitions. Wier et al's data, between 400 Hz and 2 kHz, showed relatively level frequency DLs; Moore's data were fairly level over frequencies of 250 Hz to 2 kHz for durations of 25 - 200 ms. The current data, where level over frequency, (i.e. at 40 ms) were about three times worse than Moore's data at the nearest comparable stimulus duration (50 ms).

Finally, the three sets of data were transformed in both frequency and time domains, being temporally integrated, and also differentiated with respect to critical bandwidth. This was done in an attempt to see if a combination of the two normalisation methods would 'straighten out' the data and make it independent of either frequency, duration or both. This resulted in the data being expressed in ms (i.e. (Δf / CB) x d). This made Wier et al's data level between 600 and 2000 Hz and the current data still reasonably level at 40 ms but inconsistent at other durations. Moore's data were flatter with respect to frequency than when just temporally integrated, and more so with respect to duration than when expressed in terms of critical bandwidth alone.

None of the above methods completely normalised any of the experimental data for any study; when they managed to do so it was always within certain bands of frequency or duration. It did however enable one to see that the same tendencies were exhibited in the frequency discrimination and frequency transition detection data when expressed in purely frequency terms (Δf), relative frequency terms

(Δf / f) or transformed in terms of frequency and/or duration. It was suggested in Chapter 4.5.1 that related experiments, due to evidence of sensory integration, need perhaps only be performed at a single duration. With the results of the above attempts at normalisation through integration, it would not have been advisable for the current set of experiments.

5.6 Summary and Conclusions

Frequency transition detection thresholds were obtained for pure tones over a frequency range of 250 - 2320 Hz and for durations of 10 - 80 ms in conditions similar to those used for isolated vowels in the previous chapter.

The thresholds for isolated vowels when compared with pure tones showed similar psychophysical functions over both duration and frequency. The pure tone data were on average superior to formant frequency transition detection thresholds by a factor of 4.5:1, though this was highly dependent on frequency.

The pure tone thresholds were similar to frequency DLs obtained by other workers in terms of the contour both sets of data produced with respect to duration and frequency. The frequency DL data showed consistently lower thresholds than frequency transition detection thresholds under comparable conditions. It must be remembered, however, that both Moore (1973a) and Wier et al (1977) used highly trained subjects, i.e. they performed the given task for at least 20 hours before any experimental data were collected. The subjects in the current study can only be considered as partly trained in comparison.

Where stimulus durations were closest, Moore's data were on average between 2 to 3 times better for frequency DLs than the

current data on frequency transition detection. This is only a rough estimate of the difference in difficulty between the two tasks and should be viewed with caution especially when amount of training given in both studies is considered. Both sets of data are from one subject only.

Frequency discrimination and frequency transition detection do appear to be closely related abilities as is demonstrated by the comparisons made in Chapter 5.5. The excitation pattern model used in the previous chapter could be used to compare frequency discrimination and transition detection directly. However, matching Moore's and the current stimuli in terms of duration and test frequency is quite problematic. The parameters for both do not really coincide, and as we have seen, there is still a durational dependency inherent in the modelling method employed in Chapter 4.

It would appear, though, that the current method employed for the estimation of frequency transition detection thresholds gives consistent results and reliable data. It must also be remembered that the main aim of the present chapter was to produce frequency transition data for direct comparison with the vowel data of Chapter 4.

CHAPTER 6

Frequency Transition Detection With Transitions In Initial Position

6.1 Introduction

Mattingly et al (1971) performed a series of experiments where subjects had to discriminate between different rates of second formant transition in both initial and final position for the vowel /{/. They also tested subjects on isolated F2 transitions, i.e. with no initial or final steady-state segment. In comparing the discrimination of F2 transitions in stop-vowel syllables and in isolation, and in both initial and final position, they provided evidence for a speech mode of perception.

The present experiment involves the detection of F2 transitions in initial position. These were employed in order to ascertain whether similar effects to those of Mattingly et al for rate discrimination could be observed in a similarly constructed frequency transition detection task. Liberman et al (1954) found that F2 transitions in final position were less effective as cues for the nasal consonants /m,n,ŋ/ than when presented in initial position. This weakening of cues for articulation in final position transitions was also a motivation for the present study.

6.2 Methods

Stimuli were generated by the same method as that reported for isolated vowels in Chapter 2, and for sinewaves in Chapter 5. Four different stimuli were employed, these were the vowels /3/, /ɪ/ and /a/, and a 1420 Hz sinewave (F2 frequency of /3/). Rising and

falling transitions of the second formants of each of the three vowels were produced. As with the studies using transitions in final position (Chapters 4 and 5) the total stimulus duration was 200 ms with transitions within this period of 10, 20, 40 and 80 ms.

As can be seen from the results section of this chapter the task of frequency transition detection was more difficult with the transitions in initial position in comparison to the final position data collected in the previous two chapters. Consequently, rates-of-change of frequency employed were a little greater than previously, and were between 0.5 and 128 Hz/ms for the F2's of the three vowels, and between 0.0225 and 11.3 Hz/ms for the sinewaves.

Overall sound levels were maintained at the intensities used in Chapter 4 for isolated vowels (approximately 85 dB SPL) and Chapter 5 for sinewaves (75 dB SPL).

One subject (PC) was used in this experiment. Time constraints were the main reason for only one subject being used, but this was thought to be acceptable as consistency of subjects and reliability of method have been demonstrated in similar studies reported in the previous two chapters. Even if more than one subject had been employed for the present study, individual comparisons between conditions are the main area of interest.

One criticism that might be levelled at the present study is that owing to the steady-state portion of the vowel having a duration of greater than 100 ms then the task is one of frequency discrimination (see Chapter 1.4.6). The reference frequency, however, is the same in both the comparison and test stimulus conditions, the difference between successive test stimuli being the start frequency of F2. Therefore, the task remains one of frequency transition detection.

6.3 Results and Discussion 1

As for the thresholds of frequency transition detection with transitions in final position, all thresholds show a decreasing function with increasing transition duration. All the results were plotted in terms of critical bands, which does not completely normalise data of different frequencies (as demonstrated in the previous two chapters), but does align them more closely to each other.

Figs 6.1 - 6.4 illustrate initial position transitions of the three experimental vowels /3/, /i/ and /a/ and also a 1420 Hz sinewave (F2 frequency of /3/). Comparisons are made in each figure between upward and downward transitions towards the same reference frequency (which is the locus of the second formant in Figs 6.1 - 6.3).

Fig. 6.1 shows F2 transitions for the vowel /3/. A contrast analysis of the upward and downward transition data showed no significant difference between the two conditions. If the 10 ms data were omitted from the comparison the result became significant ($p<0.01$, 1-tailed; $p<0.05$, 2-tailed) that the downward transitions were easier to detect. For the vowel /i/ (Fig. 6.2) downward transitions were easier to detect ($p<0.01$, 1 or 2-tailed), and this also proved to be the case for /a/ ($p<0.001$, 1 or 2-tailed) which is illustrated in Fig. 6.3. If the 10 ms data was left out of the comparison made in Fig. 6.2 then the difference between the two conditions is no longer significant. The last of these initial upward and downward transition comparisons was for a 1420 Hz sinewave (see Fig. 6.4). In this case rising transitions were easier to detect ($p<0.001$, 1 or 2-tailed).

If one compares the above four results and figures with data

Key to Chapter 6 Figures

IP = initial position

FP = final position

U = upward transition

D = downward transition

SW = sinewave (1420 Hz)

All vowel stimuli are four-formant with transitions in their second formant

Fig. 6.1

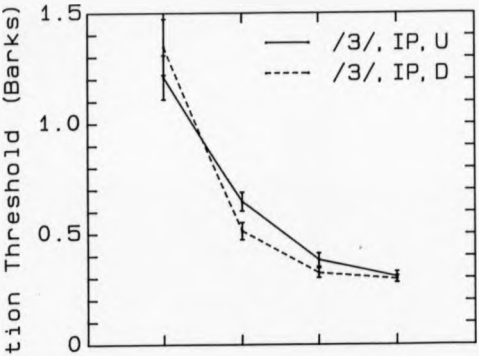

1.5
1.0
0.5
0
/3/, IP, U
/3/, IP, D

Fig. 6.2

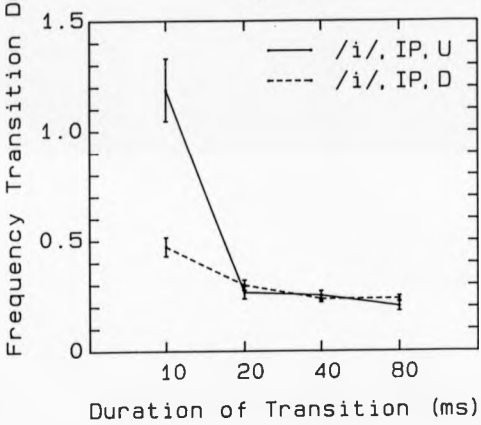

Frequency Transition Detection Threshold (Barks)
1.5
1.0
0.5
0
/i/, IP, U
/i/, IP, D
10
20
40
80
Duration of Transition (ms)

Fig. 6.3

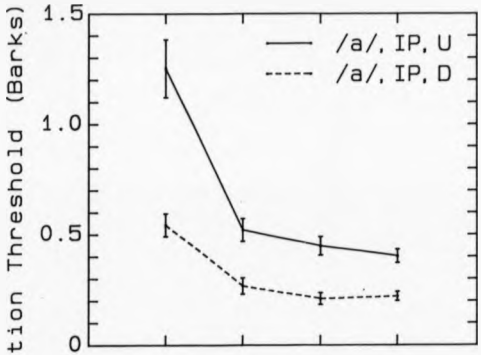
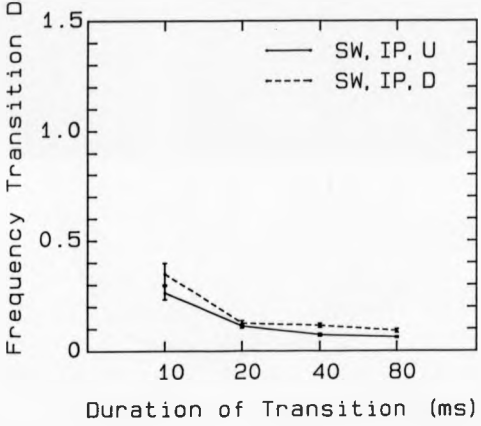

1.5
1.0
0.5
0
/a/, IP, U
/a/, IP, D

Fig. 6.4


Frequency Transition Detection Threshold (Barks)
1.5
1.0
0.5
0
SW, IP, U
SW, IP, D
10
20
40
80
Duration of Transition (ms)

from the same subject for frequency transitions in final position (see Figs 4.31 - 4.33 and Fig. 5.10) a different pattern between the upward and downward conditions emerges. For the transitions in final position there was no significant difference between upward and downward transitions of the vowel /3/ (Fig. 4.31), nor was there for a 1420 Hz sinewave (Fig. 5.10). In both cases, however, there was a slightly greater sensitivity to rising transitions. This finding is supported in the present study by the comparison between the F2 transitions of /3/ in initial position, but only if the 10 ms data is included in the comparison.

What is more interesting, though, is the result for sinewaves in initial position. Here the data for rising and falling transitions shows a consistently greater sensitivity to a rising tone sweep. It is not unusual to find a tendency in this direction in the psychoacoustic literature, for example, Collins and Cullen (1978), Nabelek (1978) and Gardner and Wilson (1979). In the present case the difference between rising and falling transitions is more pronounced than in the above studies.

One explanation for this result is that as a particular task becomes more difficult, then known differences between two conditions will be emphasised. (Mattingly et al (1971) showed that it is harder to discriminate between transition rates when the transitions are in initial rather than final position). This can be illustrated in the data of frequency transition detection for the shorter durations (i.e. 10 and 20 ms) where two stimuli of differing complexity diverge far more in threshold level (see comparisons between four- and isolated-formants in Chapter 4).

Also, Cullen and Collins (1982) found that the magnitude of the difference between rising and falling tone glides was dependent on rate-of-change of frequency. If the task is more difficult, which it

is for initial position transitions, then transitions with a larger gradient must be used to obtain thresholds at a fixed duration of transition.

In Chapter 4 the results of contrast analyses between rising and falling F2 transitions, as well as all other stimuli, were predicted on the basis of the masking effect of adjacent frequency components within (or very close to) a critical band's distance of the particular formant in motion. This viewpoint was largely justified by the obtained results and the subsequent modelling of the data. The data for transitions in initial position show a slightly different trend.

For the vowels /i/ and /a/ the downward transitions were shown to be significantly easier to detect (if, in the case of /i/, the 10 ms threshold data are included). For /i/ this is understandable if one considers the predicted masking effect of the F3 upon a rising F2 (though admittedly the opposite trend was noted for final position transitions in Fig. 4.32). One would not have predicted the same outcome for the vowel /a/, however, where the F1 should produce a masking effect on downward transitions (particularly at short durations where the frequency excursion required to obtain threshold is largest).

If one assumes that the expected result for /a/ in final position was 'correct', and perhaps less convincingly, that the prediction for /i/ in final position was incorrect due to some unforeseen factor, then the claims made by Mattingly et al would appear to have some support. They said that speech has its own mode of perception and that there is a tendency for this mode to produce categorical results (i.e. certain speech perception tasks produce results that are dependent on the cuing power of the stimulus and not solely on its spectral content). The evidence presented here is only

marginal, and a fuller discussion of the issue is presented in the following section.

Figs 6.5 - 6.20 clearly demonstrate, for both isolated vowels and pure-tones, that transitions in final position are easier to detect than those in initial position. The above numbered figures compare each of the two initial position conditions with those in final position producing four different combinations for contrast analysis for each stimulus type. It is difficult to decide whether the most appropriate comparison between initial and final position transitions should be between those in the same direction, or opposite direction but therefore similar frequency region. As there is a good case for either combination of results, it was decided to include both.

Out of 16 such comparisons between initial and final position transitions there was only one non-significant result (see Fig. 6.15 where the initial falling F2 of /a/ is compared with a falling F2 transition in final position). In every other case the result was significant ($p<0.001$, 1 or 2-tailed) that frequency transitions were easier to detect in final rather than in initial position. Removing the 10 ms data from the contrast analyses had no effect on the significance of any of the initial versus final position transition comparisons. On average, for all isolated vowel conditions and for each stimulus duration, final position transitions were easier by a factor of approximately 2:1. For the 1420 Hz sinewave condition the ratio was slightly greater at 2.5:1.

According to Elliot (1971), when the time interval between masker and signal is very short (in a nonsimultaneous masking task), then more backward than forward masking takes place. If one considers the steady-state portion of the present stimuli to be the masker, and the transition part to be the maskee, then for initial

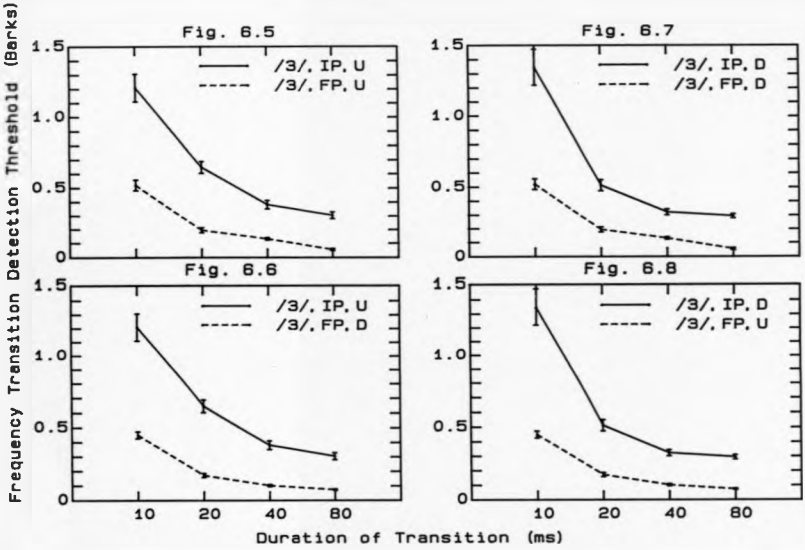


Fig. 6.5
/3/, IP, U
/3/, FP, U
Fig. 6.7
/3/, IP, D
/3/, FP, D
Fig. 6.6
/3/, IP, U
/3/, FP, D
Fig. 6.8
/3/, IP, D
/3/, FP, U
Frequency Transition Detection Threshold (Barks)
1.5
1.0
0.5
0
10
20
40
80
Duration of Transition (ms)

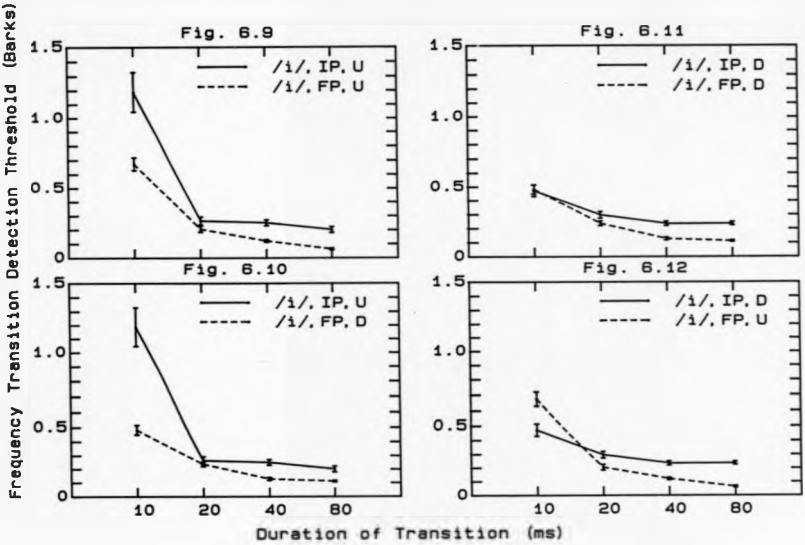


Fig. 6.9
/i/, IP, U
/i/, FP, U
Fig. 6.11
/i/, IP, D
/i/, FP, D
Fig. 6.10
/i/, IP, U
/i/, FP, D
Fig. 6.12
/i/, IP, D
/i/, FP, U
Frequency Transition Detection Threshold (Barks)
1.5
1.0
0.5
0
10
20
40
80
Duration of Transition (ms)

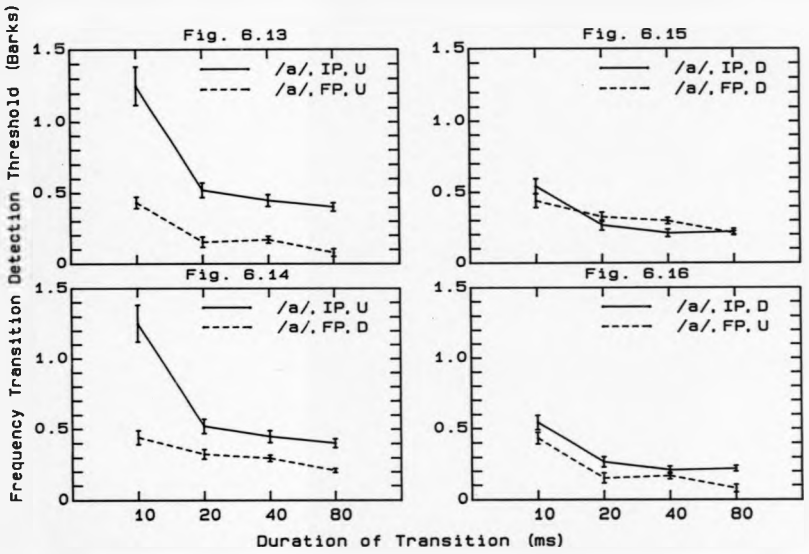


Fig. 6.13
/a/, IP, U
/a/, FP, U
Fig. 6.15
/a/, IP, D
/a/, FP, D
Fig. 6.14
/a/, IP, U
/a/, FP, D
Fig. 6.16
/a/, IP, D
/a/, FP, U
Frequency Transition Detection Threshold (Barks)
1.5
1.0
0.5
0
10
20
40
80
Duration of Transition (ms)

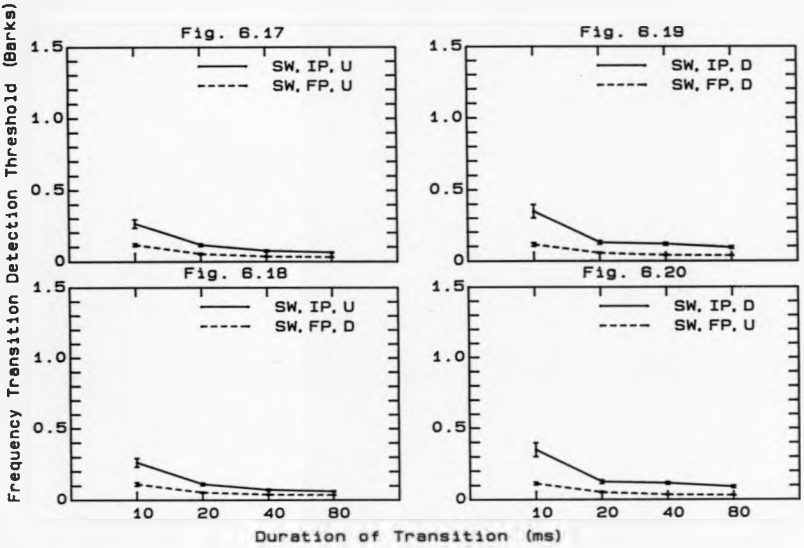


Fig. 6.17
SW, IP, U
SW, FP, U
Fig. 6.19
SW, IP, D
SW, FP, D
Fig. 6.18
SW, IP, U
SW, FP, D
Fig. 6.20
SW, IP, D
SW, FP, U
Frequency Transition Detection Threshold (Barks)
1.5
1.0
0.5
0
10
20
40
80
Duration of Transition (ms)

position transitions, backward masking occurs. As the masker and maskee are adjacent to each other then the time interval between the two is the shortest possible. Therefore, one possible explanation of the difference in thresholds between frequency transitions in initial and final position is that backward masking is stronger than forward masking (when no gap is present).

Darwin (1971) also found greater backward than forward masking when presenting listeners with consonant-vowel syllables and contralateral asynchronous (± 60 ms) maskers. In this case the maskers used were either another CV syllable, vowels or (as Darwin describes it) a non-speech steady-state timbre.

The above demonstrated difference between the detection of initial and final position frequency transitions agrees with Mattingly et al (1971) who found a similar tendency for rate discrimination tasks of F2 transitions in the vowel /(/. A recurring theme of the present series of studies is that abilities for detection and discrimination are closely related, if not being one and the same. This result also lends support to that view.

The final set of comparisons made are between vowel conditions for rising and for falling initial position F2 transitions. The data (for the purposes of contrast analysis) were expressed in terms of critical bands for between vowel comparisons. Fig. 6.21 shows rising initial F2 transitions for both /3/ and /i/. The result was significant that the rising F2 of /i/ is easier to detect than that of /3/ ($p<0.001$, 1 or 2-tailed). This contrasts sharply with the result of a comparison of the same two tasks in final position (Fig. 4.40) where the opposite trend was noted.

There was no significant difference between the rising initial F2 transitions of /3/ and /a/ (Fig. 6.22) which is a repeat of the result for final position transitions (Fig. 4.41). The comparison

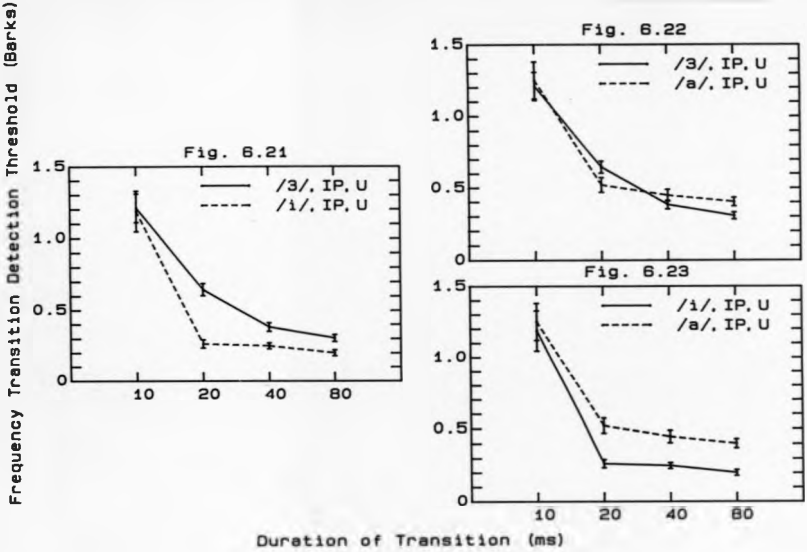


Fig. 6.21
/3/, IP, U
/i/, IP, U
Fig. 6.22
/3/, IP, U
/a/, IP, U
Fig. 6.23
/i/, IP, U
/a/, IP, U
1.5
1.0
0.5
0
10
20
40
80
Frequency Transition Detection Threshold (Barks)
Duration of Transition (ms)

between /i/ and /a/ (Fig. 6.23) produced a significant result in favour of /i/ being easier to detect ($p<0.001$, 1 or 2-tailed) which is in the opposite direction to that of the transitions for the same vowels in final position (Fig. 4.42).

Finally, Figs 6.24 - 6.26 show comparisons between vowels for falling F2 transitions in initial position. As with the rising transitions, /i/ was significantly easier to detect than /3/ ($p<0.001$, 1 or 2-tailed). The falling transitions of /a/ were also found to be easier to detect than those for /3/ ($p<0.001$,1 or 2-tailed). There was no significant difference between the vowels /i/ and /a/.

Using a psychoacoustic masking model one would have perhaps not have predicted any of the above six results, apart from the comparison between the rising F2's of /3/ and /a/ (Fig. 6.22) which produced the same result in final position. Where psychoacoustic considerations largely appeared to determine which vowel condition was easier for F2 transitions in final position, the opposite seems to have occurred here. This would also appear to lend support to the arguments that initial transitions are more powerful acoustic cues (Liberman et al, 1954) and that a speech mode of perception is in operation and is producing categorical responses for transitions in initial position.

6.4 Discussion 2 and Conclusions

First of all, it must be remembered that all the vowel stimuli referred to in the present and previous chapters employed a frequency transition in one formant alone. In order to produce the percept of a stop consonant (followed by a vowel) there needs to be a transition in at least two formants. Delattre et al (1955) showed that a strong

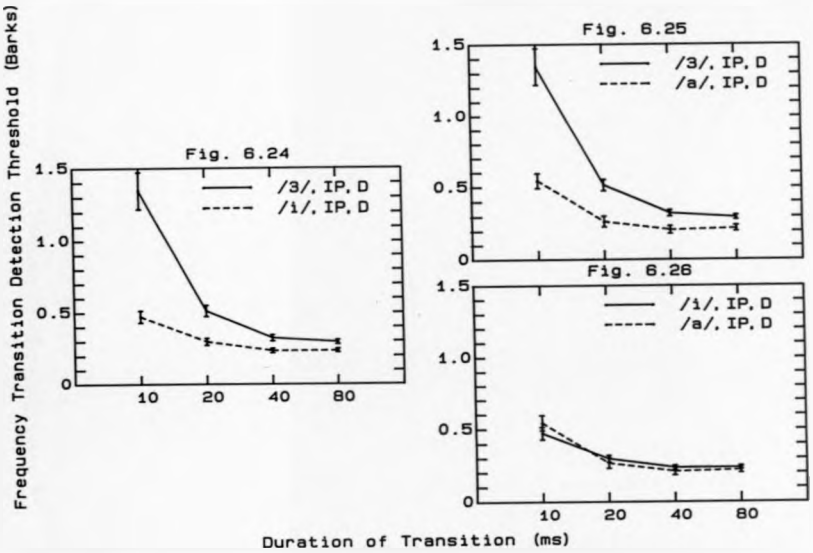


Fig. 6.24
/3/, IP, D
/i/, IP, D
Fig. 6.25
/3/, IP, D
/a/, IP, D
Fig. 6.26
/i/, IP, D
/a/, IP, D
Frequency Transition Detection Threshold (Barks)
1.5
1.0
0.5
0
10
20
40
80
Duration of Transition (ms)

voiced stop consonant could only be produced by using an F1 transition which was (comparatively) low in frequency. This is not denying that F2 transitions are important for the perception of stop consonants, as they are absolutely necessary in conjunction with the presence of an F1 transition to produce a consonantal effect. What emerges from this is that a second formant transition, in a vowel context, does not produce a strong (or full) percept of a stop consonant be the transition in initial or final position.

It is not possible to look at the covarying effect of formant transitions in vowels without first investigating sensitivities to transitions in single formants. So, at the expense of providing a listener with a strong percept of a consonant (before or after a steady-state vowel) the current studies were undertaken. Despite the artificial nature of a single formant frequency transition in a speech context (and therefore absence of a strong acoustic cue) categorical tendencies have still been in evidence for initial position transitions.

The percepts produced by the initial transitions have subjectively sounded "impulsive" or "click-like", but not completely like stop consonants. Liberman et al (1958) progressively eliminated the F1 transition in certain stop consonants leaving a silent gap in its place. The main effect of this was to convert voiced stops into voiceless ones, and was largely independent of the vowel used. There was some F3 transition as well as F2, nevertheless some sort of stop consonant preceding the vowel was still perceived.

Mattingly et al (1971) shed some useful light on why final position frequency transitions are harder to discriminate (and therefore detect). Following on from Brady et al (1961), who found that subjects made pitch matches close to the final frequency of a

rising or falling frequency sweep, they suggested the following:- "...it seems plausible to infer that for some reason subjects find it easier to estimate the final frequency of a chirp than its frequency at some earlier moment. If so, we should expect to find...that a discrimination task in which the stimuli differed most in their final frequencies and not at all in their initial frequencies would be easier than a task for which the reverse was true." ("Chirps" are described as an isolated F2 transition with no pre or post steady-state).

They went on to say that "...if we make the assumption that subjects are comparing slopes in the case of forward as well as backward chirps...(then) conceivably, a subject might estimate the slope directly." If this is true then there should be no difference between thresholds for detecting transitions in initial or final position as the task would appear to be the same (if one ignores greater effects of backward than forward masking for a multi-component stimulus). The alternative hypothesis put forward by the authors was that a subject may estimate the stimulus frequency at two successive time intervals and then compute the slope (presumably by temporal integration). The computation involved is not identical for transitions with an initial or final steady-state period as it is for transitions alone.

In final position Δt can be chosen freely (if the steady-state is presumably long enough to produce pitch memory) the frequency being already known at time t. For an initial position transition there is no prior reference, and Δt possibly cannot be chosen until the frequency of the transition has ceased to vary. The crucial point to the above argument is that transitions in final position are acoustic events where the listener moves from a known reference to an unknown target, whereas for initial position transitions both start

and destination frequency are unknown prior to the completion of the event.

In conclusion, one can say that there is clear evidence for final position frequency transitions being easier than those in initial position to detect. Also, that even with a comparatively weak acoustic cue, i.e. only a single formant transition, different patterns of responses occur for initial and final position transitions.

The evidence in the current study appears to support a psychoacoustical masking model for final position transitions, and some other perceptual model for those in initial position. Whether this is attributable to a speech mode of perception, which is brought into operation when acoustic cues are more in evidence, is difficult to say but the results of comparisons between the data from the present study and that of Chapters 4 and 5, exhibit a marked contrast with each other.

CHAPTER 7

Intensity Transition Detection and Discrimination

7.1 Introduction

The present study was designed as a parallel to those investigating frequency transition detection (Chapters 4 - 6). Specifically, a transition in the overall intensity of a vowel sound was investigated for its dependence on transition duration. One intention of the study was to see if thresholds exhibited the same decreasing function with increasing transition duration as was consistently found for frequency transitions.

7.2 Methods

Two isolated vowels were used in the present study (/3/ and /i/). The level of the steady-state portion was fixed at 80 dB SPL. This measurement was calibrated by a Bruel and Kjaer 2010 Heterodyne Analyzer used in conjunction with a 4134 1/2" microphone which was fitted inside an artificial ear (B & K). (One of the speakers from the headphones in the sound-proofed booth was clamped to the artificial ear at a pressure of 0.5 kg).

As before, transition durations used were 10, 20, 40 and 80 ms within a total stimulus duration of 200 ms, and were all in final position (with respect to the steady-state portion of the test, i.e. dynamic, stimuli). This duration has been used by other workers for intensity discrimination tasks (e.g. Viemeister, 1974), and was therefore considered to be suitable for the present task.

The construction of intensity transitions followed the same

principles as the generation of frequency transitions for previous experiments, i.e. one of linear interpolation between the reference and target value over a set period of time. Specifically, the increase in level was calculated by the equation:-

$$dB_{increment} = 20\log_{10}(A_{inc} / A_{ref})$$

where

A_{inc} = the increased amplitude of the stimulus

A_{ref} = the amplitude of the reference stimulus.

The amplitude units in this case correspond to the resolution of the digital signal, i.e. the maximum positive and negative integer values on a linear number scale (with arbitrary units).

Steady-state versions of the level-increased stimuli were produced and tested for accuracy with the Bruel and Kjaer apparatus detailed above. The above method of incrementing intensity proved to be considerably accurate. It was possible to alter the signal levels in this way due to the isolated vowels having symmetrical waveforms (in amplitude space) around zero; i.e. the sum of the digitised amplitude values within one glottal cycle was always zero. A normalising procedure in the speech synthesis software eliminated any DC-shift in the waveform.

Both increasing and decreasing transitions were used for the vowel /3/. The thresholds for increasing level transitions for the vowel /i/ were also obtained. Two subjects were employed, and both took part in all the experimental conditions.

The interval ratio between levels of difficulty of the stimuli was dependent on how hard each particular experimental condition proved to be. Sets of pilot stimuli were created whose intervals were in equal decibel increments. If these proved to be too easy (on the whole) then the range of intensity increments was further restricted, and each of the six levels adjusted (on a linear dB

scale) accordingly. The opposite was done if the converse was true. The criteria for dealing with 'easy' and 'hard' stimulus sets are set out in Chapter 2.2.

Formants were not singly increased or decreased in level as this does not occur in natural speech. If one formant is increasing or decreasing in intensity it is because there is a change in energy throughout the whole speech production system, and hence other formants are changing in level also.

Written instructions to subjects are detailed in Appendix A. The design of the experiment conforms to that for frequency transition detection in all other respects.

7.3 Results and Discussion

A decreasing threshold function for increasing transition duration is observed for intensity transition detection thresholds when expressed in terms of level increment or decrement (Figs 7.1 - 7.6). This was the general trend for frequency transition detection and is perhaps expected in the light of a study by Moore and Raab (1975); they found that for various widths of bandstop noise that intensity discrimination was poorer at 10 ms than at 250 ms.

The increasing (in level) transitions, as illustrated in Figs 7.1, 7.2 and 7.5, proved to be significantly easier to detect than decreasing transitions for the vowel /3/ in both individual and pooled conditions ($p<0.001$, 1 or 2-tailed). One might expect this to be the case as one would predict difficulties in the task at short transition durations where the stimulus would be dying away due to offset gating anyway in the last 5 ms. At the longer durations of decrease, the task was in danger of becoming one of duration discrimination as the test stimulus would audibly 'die away' quicker

Fig. 7.1

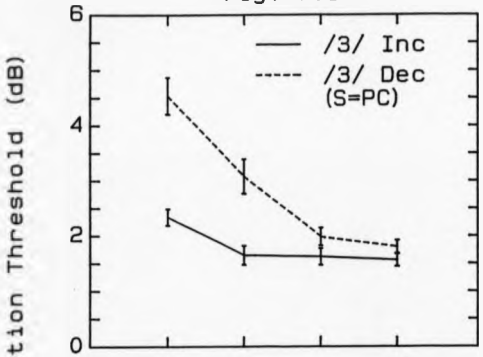
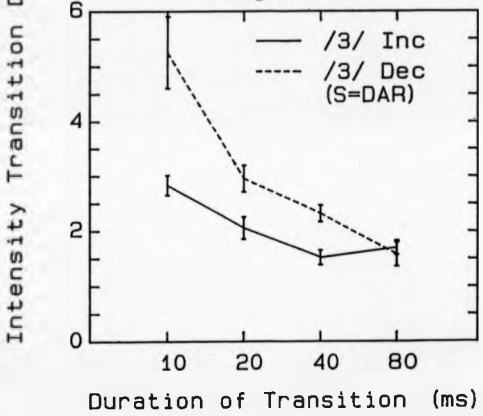


Fig. 7.2



Intensity Transition Detection Threshold (dB)

Fig. 7.3

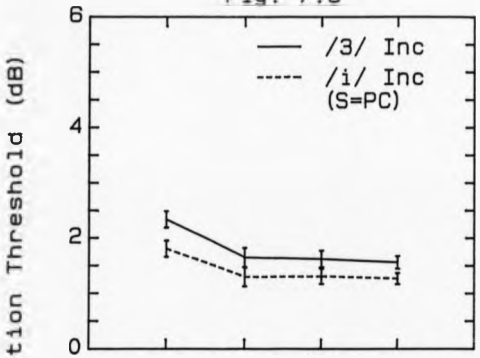

/3/ Inc
/i/ Inc
(S=PC)
6
4
2
0

Fig. 7.4

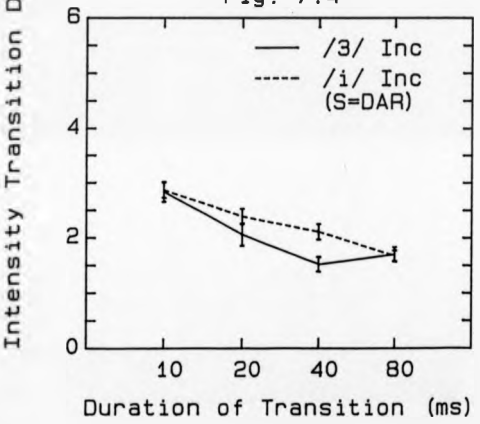

/3/ Inc
/i/ Inc
(S=DAR)
6
4
2
0
10
20
40
80
Intensity Transition Detection Threshold (dB)
Duration of Transition (ms)

Fig. 7.5

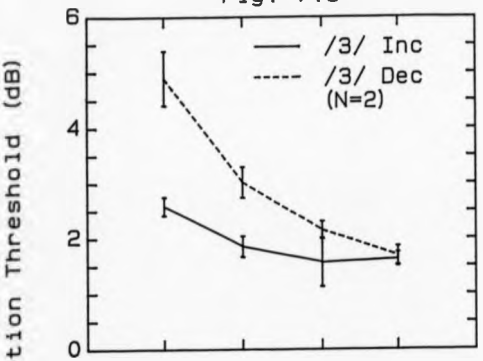
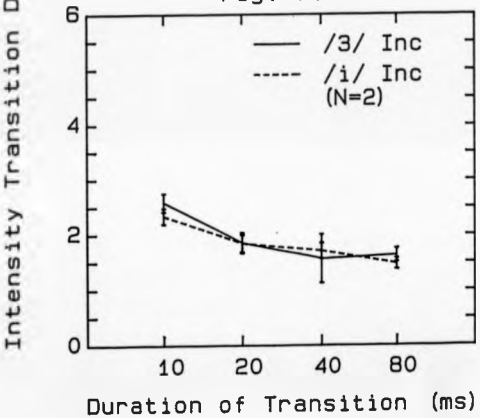

/3/ Inc
/3/ Dec
(N=2)
6
4
2
0

Fig. 7.6


/3/ Inc
/i/ Inc
(N=2)
6
4
2
0
10
20
40
80
Intensity Transition Detection Threshold (dB)
Duration of Transition (ms)

than the comparison.

Figs 7.3 and 7.4 illustrate individual comparisons between increasing intensity transitions for the vowels /3/ and /i/. Owing to equal amounts of energy being present in both sets of stimuli (they were both calibrated to a reference of 80 dB SPL and DC-normalised) one would expect there to be no difference between the two conditions. They obviously have different concentrations of energy throughout their respective spectra, but contained the same amount of energy overall.

This prediction proved correct for one subject (DAR) whose contrast analysis of the two conditions proved to be non-significant (Fig. 7.4). However, for the other subject (PC) the result was significant that increasing transitions in the vowel /i/ were easier to detect than those of the vowel /3/ ($p<0.001$, 1 or 2-tailed)(Fig. 7.3). When the two results were pooled (see Fig. 7.6) the difference was not significant as previously predicted.

One subject (PC) carried out some intensity discrimination experiments on the vowel /3/. The reference, as before, was set to 80 dB SPL; the task involved deciding which stimulus in a pair was the loudest with the test stimuli being presented at higher sound levels than the reference. The subject only successively completed trials on two stimulus durations; these were 200 and 80 ms. At shorter durations than this the stimulus no longer resembled speech and the task became governed by the differences in splatter produced by the test and reference stimuli.

The thresholds obtained were 0.61 ± 0.63 dB for 200 ms stimuli and 0.78 ± 0.11 dB for 80 ms. It is not possible to make too much from such a small sample of data, but it does appear that for the above durations intensity discrimination is superior to intensity transition detection.

Conventionally, workers have investigated intensity discrimination as a function of frequency or sensation level (e.g. Jesteadt et al, 1977). This has partially resulted in thresholds being expressed in terms of relative intensity (i.e. $\Delta I / I$). In the present case only one level was used so the above representation is not particularly appropriate. Very few studies have investigated intensity changes as a function of stimulus duration. If investigated at all, it has tended to be only a minor part of a study (e.g. Moore and Raab, 1975). Therefore, data comparisons are very difficult to make.

One possible comparison, however, is between Moore and Raab's 250 ms and 10 ms discrimination values and those of the present data for their longest and shortest transition durations. Moore and Raab found (for noise bursts) that performance for 10 ms stimuli was on average 3.8 dB worse than for intensity discrimination at 250 ms.

The present data, in comparing intensity transitions at 80 and 10 ms, show an average difference of 0.9 dB across vowels for increasing transitions. For decreasing intensity transitions (for the vowel /3/) the average difference was much higher at 3.2 dB in favour of the longer duration.

One would expect a greater difference for a more difficult task, i.e. of detecting decreasing intensity transitions. Also, Moore and Raab's data show a far greater difference than the current data presumably due to using a duration of 250 ms for the longer condition versus 80 ms in the present case.

As mentioned in Chapter 1.5, Weber's law (i.e. $\Delta I / I$ is constant) appears to hold for noise but not for pure-tone intensity discrimination. One could make the prediction that speech should also conform to Weber's law as it has energy over a relatively wide

(but ultimately limited) frequency range as does bandstop noise. This would only be assessable however if different levels were employed for the present set of vowel stimuli.

7.4 Conclusions

As few stimulus conditions were investigated it is difficult to assess the present work in comparison with major studies of intensity discrimination. What does emerge from the above study is that intensity transition detection, for isolated vowels, is dependent on stimulus duration. Thresholds largely improved with increasing transition duration in a similar relationship to that as was found for frequency transition detection.

Changing the vowel had no effect on the size of the thresholds, but transitions of decreasing intensity were harder to detect than those of increasing intensity.

CHAPTER 8

Summary and Conclusions

8.1 Summary

The purpose of this section, and indeed complete chapter, is to summarise and attempt to put into context the findings of the entire experimental program. In previous chapters the emphasis has been placed on how the obtained results compare with known psychoacoustic data; estimations of difference in orders of magnitude between the two sets of data have been provided where possible. Such an approach also forms the basis of the current chapter.

The experiments carried out on formant transition detection as a function of rate-of-change of frequency, in Chapter 3, were not particularly successful as far as interpretable data is concerned. However, they were able to point the way to a better experimental design, and served to illustrate that viewing detection of frequency transitions as a function of rate-of-change is possibly not the best way to deal with frequency transition detection.

Chapters 4 and 5 dealt with frequency transition detection in formants and sinewaves respectively. Evidence was provided for sensory integration of formant transitions in isolated vowel (and isolated formant) stimuli. The concept of masking, and its effects on frequency resolution, was able to account for the majority of the data obtained in Chapter 4. Isolated-formant stimuli produced significantly lower thresholds than for four-formant conditions. Where masking was predicted (in the four-formant conditions) the contrast was even more marked.

Expressing the threshold data in terms of percentage frequency

change proved to be unsatisfactory when attempting to compare results from different (formant) frequency conditions. As effects of masking on frequency resolution appeared to be governing the magnitude of the thresholds values, it was decided to represent the data in terms of critical bandwidth. This had a normalising effect on the data, though it did not make thresholds completely independent of frequency.

A mean threshold across all four-formant stimuli was obtained which was not much out of the range of values obtained for pure-tone frequency DL data. The values were 5.4 Hz for four-formant and between 1 and 3 Hz for pure tones at a test frequency of 1 kHz.

Zwicker's excitation pattern model sets a criterion of a difference being detectable between two signals if they vary by 1 dB somewhere in their respective excitation patterns. The pooled data from Chapter 4 were tested against this value, and when temporally averaged over the three longest transition durations showed (on the whole) good agreement with the model.

The results for detection of frequency transitions using pure tones, as reported in Chapter 5, were compared with data from two well-known frequency discrimination studies (Moore, 1973a; Wier et al, 1977). The data for each task were similar, e.g. they were both dependent on duration and described comparable contours over frequency. On average the DL data provided lower thresholds, Moore (1973a) producing values between 2 and 3 times smaller than the current data for frequency transition detection (over frequency ranges of 250 to 2320 Hz, and durations of 10 to 80 ms in the present case).

In Chapter 6 formant frequency transition detection thresholds were obtained for isolated vowels with transitions in initial position (all previous experiments having them in final position).

Without exception initial position transitions were found to be more difficult than those in final position. This was true also for pure tones, and therefore was not just a feature of the synthetic vowel stimuli. The finding is thought to be due to the absence of any reference in initial position stimuli, i.e. it is easier to detect formant movement if a steady-state portion of speech has preceded the transition in frequency.

Due to downward transitions in isolated vowels being consistently easier to detect for initial position transitions (regardless of formant proximity) it was suggested that a speech mode of perception may be in operation when speech cues are more in evidence. An initial formant transition is a more powerful acoustic cue in speech than a final one. Marslen-Wilson (1984) argues that, in speech perception, the beginnings of words are far more important than their endings for correct identification. Marslen-Wilson's 'cohort' model of spoken-word recognition states that the decision space for word identification is determined by word beginnings, the speech input at the start of a word mapping onto all lexical items that share the same initial sequence.

Finally, in Chapter 7 a parallel set of experiments were performed on intensity transition detection. The actual vowel used proved to be unimportant as far as increasing overall intensity was concerned. However, increasing intensity proved to be significantly easier to detect than decreasing transitions.

8.2 Caveats and Criticisms

There are a number of factors to take into account when assessing the value of the completed work and when attempting to put it into some sort of context.

The stimuli used, though recognised by subjects as synthetic vowel sounds, could hardly be described as natural syllabic utterances. Sreenivas and Rao (1982) point out how most stimulus signals for speech perception experiments are stationary due to the complexity of the human auditory system. The problem occurs when extending such experimental results on the basis that "...speech signals are stationary at least over short durations of the order of 30 ms." This, as they acknowledge, is just not the case in real speech where stationarity is rarely observable. As Pols and Schouten (1982) observe "In spectrograms of conversational speech one often wonders if there is anything which is not a transitional event, and if stationary, stable, segments are ever reached or do exist at all."

In order to carry out and complete such a project as this, however, as much control over stimulus conditions as possible has to be exercised. This resulted, in the present case, in elongated vowels (for a normal speech context) followed or preceded by singular formant transitions. Where speech is concerned however, coarticulation is the rule, i.e. formant transitions never really occur singly in an utterance. Also, as mentioned in Chapter 2, formant bandwidths should have been set proportional to their centre frequency rather than fixed regardless of which vowel was being synthesized. This would have meant that a particularly high and low F2, for example, (as in the vowels /i/ and /a/ respectively) would have had the same bandwidths as measured on a critical band scale.

Another issue to take into account when viewing the results is the number of subjects used for certain experimental conditions was small. Thresholds were obtained by a particularly robust method, each one being calculated from a total of 600 trials. Nevertheless, even with such safeguards it would have been better to use more subjects. Time constraints and a desire to test as many conditions

as possible necessitated such measures.

8.3 Recommendations for Further Studies

Harris et al (1958) presented subjects with a discrimination task where (in an "ABX" paradigm) a pure tone was varied in either frequency, intensity or both. The discrimination probabilities of changes in frequency and intensity in isolation were determined beforehand. They found that the addition of a second cue for discrimination always enhanced the percept that a change had taken place. This was the case even if the second cue was barely perceptible alone. In a subsequent related study Demany (1983), without criticising Harris et al's findings, argued that frequency and intensity changes cannot be detected simultaneously, i.e. separate neural information is used for the two tasks.

It would be a useful exercise, now that frequency and intensity transition detection thresholds have been obtained on the same isolated vowels (albeit at different SPL's), to perform some simultaneous detection experiments. In natural speech, formant transitions are commonly accompanied by intensity changes, which in a vowel can signal both approach to a target and stress on the particular phoneme in question. It would therefore be helpful to see if one transition aids the detection of the other even if the cues are unrelated. Harris et al showed interaction in an additive direction for all pairs of cues whether they were correlated or not.

Further work should also be done on the detection of F1 and F2 transitions in more vowels than the three already used. One ideal candidate is the vowel /u/ which has both a low F1 and F2 frequency (see Fig. 2.1). This would greatly enlarge the amount of F1 - F2 space covered and hopefully provide more evidence of masking effects.

It is desirable to investigate the effects of non-simultaneous masking on frequency transition detection thresholds. A fricative, such as /s/ or /z/, could be simulated by the filtering of white noise, and could be placed before an initial or final position formant transition in one of the previously used vowels. The effects of forward masking as a function of both silent gap (between fricative and vowel) and length of noise burst could be measured. A similar study to this has been carried out by Spenner and Urbas (1986). Two German one-syllable words were used each having a long vowel preceded by a stop consonant. They found forward masking effects in both psychophysical and neurophysiological experiments (the latter being done on anaesthetised cats).

The current implementation of Zwicker's excitation model showed encouraging enough results to be developed further. A number of problems and 'features' still need to be overcome. For example, the temporal dependency still apparent in the data analysed in Chapter 4 (and tabulated in Appendix C) needs to be eliminated in order to make the model more useful. Also, the results of Chapter 7 suggest that a 1 dB difference could well be an inadequate criterion insofar as subjects were more sensitive to increases than decreases in the overall level of a stimulus.

8.4 Final Comments

From the outset of the project the aim has been to provide more realistic psychoacoustic parameters to aid the design of automatic speech recognition devices. The work reported above is limited in its capacity to provide specific values and limits that one should include in a system for all possible phonetic combinations.

The results are useful, however, insofar as they indicate that

the problem is not all that simple. With one change of condition, for example, putting formant transitions in initial rather than final position, a whole new set of threshold values and possibly rules governing detection are produced.

The amount of work that needs to be done in this area is large, and the overall progress slow. As Pols and Schouten (1982) observed "...more work has to be done before we can understand at a psychophysical level the mechanism for detecting, discriminating, and identifying frequency sweeps occurring for instance in vocalic transitions caused by coarticulation. Meanwhile, we will go on studying in real speech the perceptual relevance of transitional information for the identification of neighbouring speech sounds. This knowledge will perhaps allow us to improve automatic speech recognition systems."

Until a lot of this work is completed models of how to recognise speech will continue to include vast areas enclosed by black boxes.

APPENDIX A

Instructions to Subjects

In addition to a verbal explanation of the psychophysical task that a subject was about to perform, some text was printed up on the screen of the monitor in the sound-proofed booth. For frequency transition detection the following text appeared:-

> "You are about to hear N pairs of synthetic vowel stimuli. Each pair contains a steady-state vowel and a vowel that contains a frequency transition in one of its formants. Your task is to decide whether the vowel that contains a transition is presented first or second in a stimulus pair. Wait until the word "READY" appears on the screen before you make your response. You must respond to all stimulus pairs. You will receive feedback after each response to tell you if your choice was correct or not. Push the switch in the "1st" direction if you think that the first vowel of the pair contained the transition, or in the "2nd" direction if you think that it was the latter of the pair."

If the subject had requested some steady-state stimuli before the start of the experiment, the following message appeared:-

> "You have just heard X examples of the steady-state stimulus. Press button on hand control to commence the experiment."

For sinewave experiments read pure-tone for synthetic vowel, and delete references to formants.

The instructions to subjects for intensity discrimination tasks were as follows:-

> "You are about to hear N pairs of synthetic vowel stimuli. Each pair contains a sound of constant loudness and an identical sound which changes in loudness towards the end of its total duration. Your task is to decide whether the sound that gets louder/softer is presented first or last in a stimulus pair. Wait until the word "READY" appears on the screen before you make your response. You must respond to all stimulus pairs. You will receive feedback after each response to tell you if your choice was correct or not. Push the switch in the "1st" direction if you think that the first sound was louder/softer, or in the "2nd" direction if you think that the latter of the pair was louder.

A subject was informed beforehand whether the test stimuli were going to get louder or softer. No stimulus trial set included increases and decreases in intensity.

APPENDIX B

Methodology of Statistical Testing

In Chapter 3, and all subsequent chapters, the method of contrast analysis was employed as an alternative to t- and Wilcoxon tests. These two methods were rejected on the grounds that they did not make use of standard deviation values supplied by the "binfit" threshold estimation computer program.

Contrast analysis scores over many other statistical tests insofar as it is a focused test. That is to say, contrasts address precise questions about obtained data, and are able to evaluate specific predictions by comparisons with such data. Many statistical tests can be referred to as 'omnibus' tests as they are "...tests of significance that address diffuse (or unfocused) questions." (Rosenthal and Rosnow, 1985).

The rationale behind the test is detailed as follows:-

The hypothesis, which is to be refuted, is that the mean difference between the two sets of scores in a related task is zero. i.e.

$$\Sigma d_i \sim N(0,\sigma^2)$$

where i = the index of a difference between scores, and

σ^2 = the variance of the quantity Σd_i.

The variance is estimated by

$$\Sigma(\sigma^2_{Ai} + \sigma^2_{Bi})$$

where A and B are the two conditions to be compared, and σ is the standard deviation of a particular score.

It follows that

$$\Sigma_i d_i \; / \; [\Sigma(\sigma^2_{Ai} + \sigma^2_{Bi})]^{1/2} \sim N(0,1)$$

According to Bischof et al (1985) in contrasts for differences between conditions (under the hypothesis that there is no effect), the quantity produced from the above expression should be distributed as the standard normal variable z.

For frequency transition detection experiments the levels of difficulty (i.e. rates-of-change of frequency) were converted to levels separated by a linear number scale. This was done in order to keep the chi-squared values produced by the "binfit" program at a low level (in comparison with the obtained mean), and hence make the threshold estimation more reliable.

One consequence of reconverting the obtained threshold and standard deviation values to their original scaling is that the standard deviation (due to non-linear expansion) becomes slightly skew. After reconversion the mean is not symmetrically placed between the upper and lower standard deviation values. This has no effect on the significance scores from contrast analysis but could affect tabulated and graphical representations of the data. In order to ascertain the full effect of this the mean, mean + standard deviation and mean - standard deviation values were transformed separately from the linear "binfit" scale to the ratio scale used for representing the data. They were then tested (for extreme examples of experimental data conditions and scores) to see how skew the functions were. At no time was the deviation great enough to have to produce two values of standard deviation for each mean value.

Where reference conditions were identical, e.g. for four- versus isolated-formant frequency transitions, contrast analysis was

performed on the "binfit" linear number scale values. Where reference conditions were different, e.g. between different formants of the same vowel or between different vowels, the analysis was carried out using the original experimental scaling. This is because rate-of-change of frequency expressed in Hz/ms (for a frequency transition detection task) is meaningless unless the frequency region the varying formants occupy is taken into account.

APPENDIX C

Supplementary Tables

Table 1: dB differences for the rising F1 of /3/

channel	frequency	80ms	40ms	20ms	average
1	111	-0.1041	-0.0085	-0.0041	-0.0389
2	152	0.0743	-0.0499	-0.0542	-0.0099
3	196	-0.1051	-0.0546	-0.0523	-0.0707
4	246	-0.0895	-0.0891	-0.0770	-0.0852
5	300	-0.1517	-0.1379	-0.1288	-0.1395
6	360	-0.3478	-0.3480	-0.4139	-0.3699
7	427	-0.4407	-0.4086	-0.3928	-0.4140
8	500	-0.6343	-0.6212	-0.6373	-0.6309 *
9	581	0.5338	0.2730	-0.2070	0.1999
10	670	0.8239	0.9093	1.2903	1.0078 *
11	769	0.6771	0.6592	0.7872	0.7078 *
12	878	0.3773	0.4041	0.6070	0.4628
13	1000	0.2276	0.2730	0.5378	0.3461
14	1135	0.1144	0.1294	0.2594	0.1677
15	1286	0.0261	0.0284	0.0498	0.0348
16	1454	0.0014	0.0015	0.0042	0.0024
17	1642	-0.0191	-0.0242	-0.0556	-0.0330
18	1852	-0.0463	-0.0587	-0.1212	-0.0754
19	2089	-0.0435	-0.0603	-0.1259	-0.0766
20	2356	-0.0172	-0.0238	-0.0503	-0.0304
21	2657	0.0055	0.0033	0.0011	0.0033
22	2999	0.0108	0.0126	0.0302	0.0179
23	3389	0.0005	-0.0037	0.0024	-0.0003
24	3836	-0.0053	-0.0141	-0.0222	-0.0139

Table 2: dB differences for the rising F2 of /3/

channel	frequency	80ms	40ms	20ms	average
1	111	0.0023	0.0027	0.0014	0.0021
2	152	0.0034	0.0035	0.0055	0.0041
3	196	0.0025	0.0014	0.0015	0.0018
4	246	0.0030	0.0027	0.0037	0.0031
5	300	0.0059	0.0050	0.0031	0.0047
6	360	0.0077	0.0079	0.0074	0.0077
7	427	0.0067	0.0063	0.0060	0.0063
8	500	0.0044	0.0036	0.0028	0.0036
9	581	-0.0005	-0.0009	-0.0012	-0.0009
10	670	-0.0029	-0.0023	-0.0003	-0.0018
11	769	-0.0158	-0.0176	-0.0178	-0.0171
12	878	-0.0498	-0.0640	-0.0736	-0.0625
13	1000	-0.1189	-0.1570	-0.2001	-0.1587
14	1135	-0.2882	-0.3690	-0.4490	-0.3687
15	1286	-0.6314	-0.7060	-0.6525	-0.6633 *
16	1454	-0.1673	-0.1111	0.0233	-0.0850
17	1642	0.6927	0.9329	1.1720	0.9325 *
18	1852	0.5042	0.7395	1.0725	0.7721 *
19	2089	0.2522	0.3615	0.5233	0.3790
20	2356	0.0481	0.0755	0.1139	0.0792
21	2657	0.0004	0.0012	-0.0007	0.0003
22	2999	-0.0324	-0.0395	-0.0609	-0.0443
23	3389	-0.0123	-0.0165	-0.0207	-0.0165
24	3836	0.0047	0.0080	0.0098	0.0075

Table 3: dB differences for the rising isolated F2 of /3/

channel	frequency	80ms	40ms	20ms	average
1	111	-0.0009	-0.0004	0.0000	-0.0004
2	152	-0.0014	-0.0019	-0.0013	-0.0015
3	196	-0.0024	-0.0024	-0.0019	-0.0022
4	246	-0.0035	-0.0036	-0.0021	-0.0031
5	300	-0.0086	-0.0071	-0.0040	-0.0066
6	360	-0.0106	-0.0094	-0.0095	-0.0098
7	427	-0.0122	-0.0115	-0.0102	-0.0113
8	500	-0.0226	-0.0239	-0.0199	-0.0221
9	581	-0.0312	-0.0319	-0.0310	-0.0314
10	670	-0.0442	-0.0461	-0.0512	-0.0472
11	769	-0.0625	-0.0714	-0.0890	-0.0743
12	878	-0.0994	-0.1114	-0.1451	-0.1186
13	1000	-0.1671	-0.2001	-0.2753	-0.2142
14	1135	-0.3186	-0.3873	-0.5176	-0.4078
15	1286	-0.6001	-0.6326	-0.6417	-0.6248 *
16	1454	-0.1649	-0.1067	-0.0015	-0.0910
17	1642	0.6177	0.7829	1.0506	0.8171 *
18	1852	0.4990	0.6912	1.0596	0.7499 *
19	2089	0.3593	0.4820	0.7469	0.5294 *
20	2356	0.2856	0.3915	0.5821	0.4197
21	2657	0.2321	0.3255	0.4873	0.3483
22	2999	0.1986	0.2850	0.4398	0.3078
23	3389	0.2180	0.3080	0.4376	0.3212
24	3836	0.2107	0.2476	0.3874	0.2819

Table 4: dB differences for the rising F1 of /i/

channel	frequency	80ms	40ms	20ms	average
1	111	-0.0385	-0.0280	-0.0160	-0.0275
2	152	-0.2423	-0.3576	-0.3108	-0.3036
3	196	-0.2168	-0.2070	-0.0887	-0.1708
4	246	-0.0996	-0.0232	0.2024	0.0265
5	300	0.5939	0.6321	0.3172	0.5144 *
6	360	0.6362	0.8848	0.7183	0.7464 *
7	427	0.4654	0.6176	0.5245	0.5358 *
8	500	0.3793	0.5375	0.5603	0.4924
9	581	0.3136	0.4817	0.5560	0.4504
10	670	0.2656	0.4219	0.5567	0.4147
11	769	0.2627	0.4192	0.5034	0.3951
12	878	0.1706	0.3035	0.3850	0.2864
13	1000	0.1398	0.2585	0.3743	0.2575
14	1135	0.1194	0.2102	0.3245	0.2180
15	1286	0.1662	0.2514	0.3349	0.2508
16	1454	0.1029	0.1684	0.2935	0.1883
17	1642	0.0850	0.1198	0.1957	0.1335
18	1852	0.0104	0.0328	0.0409	0.0280
19	2089	0.0035	0.0070	0.0031	0.0045
20	2356	-0.0013	-0.0007	-0.0017	-0.0012
21	2657	-0.0029	-0.0017	-0.0024	-0.0023
22	2999	-0.0024	-0.0016	-0.0021	-0.0020
23	3389	-0.0046	-0.0017	-0.0021	-0.0028
24	3836	-0.0041	-0.0029	-0.0035	-0.0035

Table 5: dB differences for the rising F2 of /ɨ/

channel	frequency	80ms	40ms	20ms	average
1	111	-0.0001	-0.0002	0.0002	0.0000
2	152	0.0001	0.0002	0.0008	0.0004
3	196	0.0002	0.0003	0.0002	0.0002
4	246	0.0000	0.0001	0.0000	0.0000
5	300	-0.0003	-0.0006	-0.0007	-0.0005
6	360	-0.0001	-0.0002	-0.0005	-0.0003
7	427	-0.0001	-0.0002	-0.0019	-0.0007
8	500	-0.0010	-0.0019	-0.0083	-0.0037
9	581	-0.0082	-0.0163	-0.0148	-0.0131
10	670	-0.0238	-0.0472	-0.0239	-0.0316
11	769	-0.0110	-0.0223	-0.0300	-0.0211
12	878	-0.0545	-0.1108	-0.1061	-0.0905
13	1000	-0.0814	-0.1651	-0.1773	-0.1413
14	1135	-0.0920	-0.1903	-0.1586	-0.1470
15	1286	-0.0783	-0.1517	-0.1735	-0.1345
16	1454	-0.1808	-0.3732	-0.4683	-0.3408
17	1642	-0.3362	-0.7129	-0.8340	-0.6277 *
18	1852	-0.6632	-1.4698	-1.7584	-1.2971 *
19	2089	-0.8370	-1.9855	-2.1498	-1.6574 *
20	2356	-0.1342	-0.2863	-0.2864	-0.2356
21	2657	0.0585	0.1224	0.1137	0.0982
22	2999	-0.0283	-0.0592	-0.0738	-0.0538
23	3389	-0.0092	-0.0195	-0.0234	-0.0174
24	3836	0.0053	0.0113	0.0127	0.0098

Table 6: dB differences for the rising isolated F2 of /i/

channel	frequency	80ms	40ms	20ms	average
1	111	0.0005	0.0003	0.0005	0.0004
2	152	0.0005	0.0009	0.0006	0.0007
3	196	0.0004	0.0012	0.0003	0.0006
4	246	0.0004	0.0019	0.0001	0.0008
5	300	-0.0006	-0.0012	-0.0013	-0.0010
6	360	-0.0022	-0.0026	-0.0025	-0.0024
7	427	-0.0030	-0.0004	-0.0003	-0.0012
8	500	-0.0101	-0.0095	-0.0048	-0.0081
9	581	-0.0037	-0.0028	-0.0036	-0.0034
10	670	-0.0095	-0.0119	-0.0072	-0.0095
11	769	-0.0186	-0.0162	-0.0194	-0.0181
12	878	-0.0274	-0.0256	-0.0347	-0.0292
13	1000	-0.0376	-0.0380	-0.0563	-0.0440
14	1135	-0.0482	-0.0417	-0.0642	-0.0514
15	1286	-0.0722	-0.0684	-0.0960	-0.0789
16	1454	-0.1007	-0.0932	-0.1517	-0.1152
17	1642	-0.1612	-0.1527	-0.2563	-0.1901
18	1852	-0.3665	-0.3394	-0.5591	-0.4217
19	2089	-0.8638	-0.7108	-1.0299	-0.8682 *
20	2356	-0.3093	-0.2176	-0.1381	-0.2217
21	2657	0.5853	0.5184	0.9755	0.6931 *
22	2999	0.4569	0.4275	0.8063	0.5636 *
23	3389	0.3295	0.3092	0.5526	0.3971
24	3836	0.2624	0.2477	0.4395	0.3165

Table 7: dB differences for the rising F1 of /a/

channel	frequency	80ms	40ms	20ms	average
1	111	0.0067	-0.0041	-0.0026	0.0000
2	152	-0.0273	-0.0390	-0.0356	-0.0340
3	196	-0.0408	-0.0453	-0.0409	-0.0423
4	246	-0.0604	-0.0667	-0.0452	-0.0574
5	300	-0.1129	-0.1188	-0.1017	-0.1111
6	360	-0.1994	-0.2142	-0.2212	-0.2116
7	427	-0.2275	-0.2284	-0.2236	-0.2265
8	500	-0.3572	-0.3565	-0.3533	-0.3557
9	581	-0.5723	-0.5727	-0.6009	-0.5820 *
10	670	-0.7685	-0.7833	-0.7776	-0.7765 *
11	769	-0.2027	-0.3373	-0.5882	-0.3761
12	878	0.2747	0.3471	0.5099	0.3772
13	1000	0.1534	0.1326	0.1185	0.1348
14	1135	-0.0253	-0.0513	-0.0896	-0.0554
15	1286	-0.1005	-0.1508	-0.2758	-0.1757
16	1454	-0.1838	-0.2640	-0.4894	-0.3124
17	1642	-0.2318	-0.3224	-0.5777	-0.3773
18	1852	-0.1947	-0.2845	-0.4968	-0.3253
19	2089	-0.0760	-0.1214	-0.2107	-0.1360
20	2356	0.0018	-0.0144	-0.0242	-0.0123
21	2657	0.0245	0.0108	0.0198	0.0184
22	2999	0.0390	0.0384	0.0580	0.0451
23	3389	0.0224	0.0124	0.0160	0.0169
24	3836	0.0056	-0.0125	-0.0217	-0.0095

Table 8: dB differences for the rising F2 of /a/

channel	frequency	80ms	40ms	20ms	average
1	111	-0.5509	-0.7298	-0.6514	-0.6440 *
2	152	-0.5479	-0.7350	-0.6546	-0.6458 *
3	196	-0.5420	-0.7347	-0.6572	-0.6446 *
4	246	-0.5344	-0.7254	-0.6557	-0.6385 *
5	300	-0.5222	-0.7121	-0.6500	-0.6281 *
6	360	-0.4928	-0.6772	-0.6070	-0.5923 *
7	427	-0.4884	-0.6683	-0.6050	-0.5872 *
8	500	-0.4735	-0.6667	-0.5967	-0.5790 *
9	581	-0.4672	-0.6569	-0.5941	-0.5727 *
10	670	-0.5014	-0.6865	-0.6183	-0.6021 *
11	769	-0.5579	-0.7404	-0.6551	-0.6511 *
12	878	-0.6608	-0.8283	-0.7066	-0.7319 *
13	1000	-0.7978	-0.9741	-0.8167	-0.8629 *
14	1135	0.6106	0.3295	0.0933	0.3445
15	1286	0.6623	0.5524	0.5840	0.5996 *
16	1454	0.3879	0.3735	0.5655	0.4423
17	1642	0.0976	0.0187	0.1994	0.1052
18	1852	-0.1399	-0.2666	-0.0915	-0.1660
19	2089	-0.3870	-0.5350	-0.3983	-0.4401
20	2356	-0.5392	-0.7165	-0.6162	-0.6240 *
21	2657	-0.5736	-0.7567	-0.6754	-0.6686 *
22	2999	-0.6116	-0.8020	-0.7398	-0.7178 *
23	3389	-0.5639	-0.7422	-0.6770	-0.6610 *
24	3836	-0.5534	-0.7343	-0.6543	-0.6473 *

Table 9: Level Difference Matrix for rising F1 of /3/

20 ms transition

		time (ms) from start of transition		
channel	frequency	0	5	10
1	111	0.00	-0.01	-0.01
2	152	0.00	-0.07	-0.11
3	196	-0.01	-0.05	-0.10
4	246	0.02	-0.10	-0.17
5	300	-0.02	-0.17	-0.24
6	360	-0.11	-0.53	-0.75
7	427	0.00	-0.74	-0.85
8	500	0.02	-0.71	-1.43
9	581	-0.04	-0.25	-0.38
10	670	0.16	1.62	2.21
11	769	0.25	1.51	1.34
12	878	0.49	0.67	0.72
13	1000	0.24	0.33	0.82
14	1135	0.13	0.16	0.39
15	1286	0.02	0.04	0.08
16	1454	0.00	0.00	0.01
17	1642	-0.03	-0.03	-0.09
18	1852	-0.05	-0.05	-0.20
19	2089	-0.06	-0.06	-0.20
20	2356	-0.03	-0.03	-0.08
21	2657	0.00	0.00	0.01
22	2999	0.01	0.01	0.05
23	3389	0.00	0.00	0.01
24	3836	-0.01	-0.01	-0.04

Table 10: Level Difference Matrix for rising F1 of /3/

40 ms transition

channel	frequency	time (ms) from start of transition						
		0	5	10	15	20	25	30
1	111	0.00	0.00	-0.01	-0.01	-0.01	-0.02	-0.02
2	152	0.00	-0.02	-0.03	-0.05	-0.07	-0.09	-0.10
3	196	0.00	-0.01	-0.02	-0.05	-0.07	-0.10	-0.12
4	246	0.00	-0.01	-0.03	-0.09	-0.12	-0.17	-0.21
5	300	-0.01	-0.04	-0.06	-0.13	-0.18	-0.28	-0.31
6	360	-0.02	-0.12	-0.20	-0.40	-0.49	-0.67	-0.71
7	427	0.00	-0.20	-0.22	-0.52	-0.57	-0.84	-0.90
8	500	0.01	-0.16	-0.39	-0.63	-0.87	-1.15	-1.37
9	581	-0.01	-0.03	0.04	0.27	0.41	0.61	0.64
10	670	0.03	0.42	0.59	1.03	1.20	1.45	1.67
11	769	0.07	0.34	0.38	0.86	0.85	1.37	1.33
12	878	0.11	0.15	0.35	0.43	0.62	0.71	0.53
13	1000	0.05	0.07	0.18	0.22	0.34	0.38	0.51
14	1135	0.03	0.04	0.09	0.10	0.17	0.18	0.23
15	1286	0.01	0.01	0.02	0.03	0.04	0.04	0.05
16	1454	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	1642	0.00	0.00	-0.02	-0.02	-0.03	-0.03	-0.05
18	1852	-0.01	-0.02	-0.04	-0.04	-0.07	-0.07	-0.11
19	2089	-0.01	-0.01	-0.05	-0.05	-0.08	-0.08	-0.11
20	2356	-0.01	-0.01	-0.01	-0.01	-0.03	-0.03	-0.05
21	2657	0.00	0.01	0.01	0.00	0.00	0.00	0.00
22	2999	0.00	0.00	0.01	0.01	0.01	0.01	0.03
23	3389	0.00	-0.01	-0.01	-0.01	0.00	0.00	0.01
24	3836	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.03

Table 11: Level Difference Matrix for rising F1 of /3/

80 ms transition

time (ms) from start of transition

ch	freq	15	20	25	30	35	40	45	50	55	60	65	70
1	111	-.02	.00	.00	-.01	-.01	-.01	-.02	-.02	-.02	-.02	-.03	-.03
2	152	-.04	-.01	-.05	-.06	-.07	-.08	-.09	-.09	-.09	-.09	-.08	-.08
3	196	-.03	-.04	-.05	-.06	-.08	-.10	-.11	-.12	-.13	-.13	-.13	-.13
4	246	-.03	-.05	-.08	-.10	-.11	-.13	-.15	-.17	-.17	-.17	-.17	-.17
5	300	-.06	-.08	-.12	-.14	-.19	-.21	-.25	-.27	-.29	-.30	-.31	-.30
6	360	-.17	-.21	-.29	-.33	-.41	-.46	-.53	-.57	-.61	-.61	-.61	-.58
7	427	-.23	-.26	-.39	-.41	-.54	-.57	-.69	-.72	-.78	-.79	-.79	-.81
8	500	-.26	-.39	-.49	-.61	-.72	-.82	-.94	-1.04	-1.06	-1.10	-1.11	-1.13
9	581	.15	.25	.41	.49	.62	.69	.78	.82	.90	.94	.98	.97
10	670	.49	.57	.77	.85	1.00	1.08	1.19	1.25	1.19	1.18	1.15	1.21
11	769	.38	.40	.61	.61	.84	.83	1.07	1.04	1.13	1.12	1.12	1.14
12	878	.18	.27	.30	.39	.43	.51	.56	.59	.62	.62	.62	.40
13	1000	.10	.16	.18	.23	.25	.29	.31	.34	.35	.34	.34	.35
14	1135	.06	.09	.09	.12	.12	.14	.15	.17	.17	.17	.17	.16
15	1286	.01	.02	.02	.03	.03	.04	.04	.04	.04	.04	.04	.04
16	1454	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
17	1642	.00	-.01	-.01	-.02	-.02	-.03	-.03	-.03	-.03	-.03	-.03	-.03
18	1852	-.02	-.03	-.03	-.05	-.05	-.05	-.05	-.07	-.07	-.07	-.07	-.08
19	2089	.00	-.03	-.03	-.04	-.04	-.04	-.04	-.07	-.07	-.08	-.08	-.08
20	2356	.00	-.02	-.02	-.01	-.01	-.03	-.03	-.02	-.02	-.03	-.03	-.03
21	2657	.01	.01	.01	.01	.00	.00	.00	.01	.01	.00	.00	.01
22	2999	.01	.00	.00	.01	.01	.01	.02	.02	.02	.02	.02	.02
23	3389	.00	-.01	-.01	.00	.00	.00	.00	.01	.01	.00	.00	.00
24	3836	.00	-.01	-.01	-.01	-.01	-.01	.00	.00	-.01	-.01	-.01	-.02

Table 12: Level Difference Matrix for rising isolated F2 of /3/

20 ms transition

channel	frequency	time (ms) from start of transition 0	5	10
1	111	0.00	0.00	0.00
2	152	0.00	0.00	0.00
3	196	0.00	0.00	0.00
4	246	0.00	0.00	0.00
5	300	0.00	0.00	-0.01
6	360	0.00	-0.01	-0.02
7	427	0.00	-0.01	-0.02
8	500	-0.01	-0.02	-0.03
9	581	-0.01	-0.03	-0.05
10	670	-0.02	-0.05	-0.09
11	769	-0.04	-0.07	-0.14
12	878	-0.07	-0.11	-0.23
13	1000	-0.13	-0.18	-0.43
14	1135	-0.35	-0.40	-0.71
15	1286	-0.21	-0.89	-1.27
16	1454	0.06	0.19	-0.08
17	1642	0.63	0.91	1.58
18	1852	0.55	0.60	1.55
19	2089	0.38	0.40	1.10
20	2356	0.30	0.30	0.88
21	2657	0.26	0.26	0.72
22	2999	0.21	0.22	0.67
23	3389	0.22	0.23	0.65
24	3836	0.27	0.27	0.50

Table 13: Level Difference Matrix for rising isolated F2 of /3/ 40 ms transition

		time (ms) from start of transition						
channel	frequency	0	5	10	15	20	25	30
1	111	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	152	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	196	0.00	0.00	0.00	0.00	0.00	0.00	-0.01
4	246	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01
5	300	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.02
6	360	0.00	0.00	0.00	-0.01	-0.02	-0.02	-0.02
7	427	0.00	0.00	-0.01	-0.01	-0.02	-0.02	-0.03
8	500	0.00	-0.01	-0.02	-0.02	-0.03	-0.04	-0.05
9	581	0.00	-0.01	-0.02	-0.03	-0.04	-0.06	-0.07
10	670	-0.01	-0.02	-0.03	-0.04	-0.05	-0.08	-0.09
11	769	-0.01	-0.02	-0.04	-0.06	-0.09	-0.11	-0.14
12	878	-0.02	-0.03	-0.07	-0.10	-0.14	-0.17	-0.21
13	1000	-0.04	-0.06	-0.14	-0.17	-0.25	-0.28	-0.38
14	1135	-0.12	-0.13	-0.33	-0.36	-0.55	-0.57	-0.59
15	1286	-0.06	-0.30	-0.41	-0.73	-0.90	-1.20	-1.44
16	1454	0.03	0.09	-0.02	-0.05	-0.17	-0.23	-0.30
17	1642	0.21	0.30	0.59	0.70	1.03	1.15	1.39
18	1852	0.18	0.20	0.50	0.52	0.84	0.85	1.21
19	2089	0.13	0.14	0.34	0.35	0.58	0.59	0.85
20	2356	0.13	0.13	0.28	0.28	0.47	0.48	0.68
21	2657	0.10	0.10	0.24	0.24	0.40	0.41	0.56
22	2999	0.09	0.09	0.21	0.21	0.36	0.36	0.49
23	3389	0.12	0.12	0.24	0.24	0.35	0.36	0.50
24	3836	0.09	0.09	0.17	0.17	0.35	0.34	0.40

Table 14: Level Difference Matrix for rising isolated F2 of /3/

80 ms transition

time (ms) from start of transition

ch	freq	15	20	25	30	35	40	45	50	55	60	65	70
1	111	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
2	152	.00	.00	.00	.00	.00	.00	.00	.00	.00	-.01	.00	.00
3	196	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	-.01	-.01
4	246	.00	.00	.00	.00	.00	.00	.00	-.01	-.01	-.01	-.01	-.01
5	300	-.01	-.01	-.01	-.01	-.01	-.01	-.02	-.01	-.01	-.01	-.01	-.01
6	360	.00	-.01	-.01	-.01	-.01	-.01	-.02	-.02	-.02	-.02	-.01	-.01
7	427	.00	-.01	-.01	-.01	-.02	-.02	-.02	-.02	-.02	-.02	-.02	-.02
8	500	-.01	-.02	-.02	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.04	-.04
9	581	-.01	-.02	-.03	-.03	-.04	-.04	-.05	-.05	-.05	-.05	-.06	-.06
10	670	-.02	-.03	-.03	-.04	-.06	-.06	-.06	-.07	-.07	-.07	-.07	-.08
11	769	-.03	-.04	-.05	-.06	-.07	-.08	-.09	-.10	-.10	-.10	-.10	-.10
12	878	-.05	-.07	-.08	-.10	-.11	-.13	-.14	-.15	-.15	-.16	-.16	-.15
13	1000	-.08	-.11	-.13	-.17	-.18	-.22	-.23	-.24	-.25	-.25	-.25	-.27
14	1135	-.15	-.24	-.25	-.34	-.35	-.43	-.45	-.49	-.50	-.50	-.50	-.39
15	1286	-.31	-.36	-.51	-.58	-.72	-.79	-.93	-.98	-1.03	-1.04	-1.04	-1.12
16	1454	.00	-.07	-.09	-.15	-.17	-.22	-.24	-.29	-.31	-.31	-.32	-.34
17	1642	.29	.42	.47	.60	.65	.80	.86	.94	.95	.95	.95	.89
18	1852	.21	.35	.36	.49	.50	.64	.65	.73	.73	.73	.73	.74
19	2089	.16	.26	.27	.35	.36	.47	.47	.51	.51	.51	.51	.54
20	2356	.13	.21	.21	.27	.27	.37	.37	.42	.42	.41	.41	.44
21	2657	.10	.15	.15	.22	.22	.32	.32	.34	.34	.35	.35	.35
22	2999	.07	.11	.11	.21	.21	.28	.28	.31	.31	.28	.28	.33
23	3389	.14	.14	.14	.23	.23	.27	.27	.33	.33	.32	.32	.34
24	3836	.10	.17	.17	.23	.23	.24	.24	.28	.29	.31	.31	.28

Table 15: Level Difference Matrix for rising F1 of /i/

20 ms transition

channel	frequency	time (ms) from start of transition 0	5	10
1	111	0.01	-0.05	-0.04
2	152	-0.01	-0.23	-0.64
3	196	0.00	-0.02	-0.18
4	246	0.01	0.19	0.38
5	300	0.01	0.14	0.61
6	360	0.02	0.89	1.31
7	427	0.12	0.61	0.90
8	500	0.19	0.53	0.89
9	581	0.20	0.45	0.88
10	670	0.16	0.41	0.90
11	769	0.22	0.33	0.79
12	878	0.12	0.20	0.62
13	1000	0.15	0.17	0.60
14	1135	0.14	0.17	0.55
15	1286	0.21	0.24	0.47
16	1454	0.17	0.17	0.40
17	1642	0.07	0.07	0.30
18	1852	0.01	0.01	0.07
19	2089	0.00	0.00	0.01
20	2356	0.00	0.00	0.00
21	2657	0.00	0.00	0.00
22	2999	0.00	0.00	0.00
23	3389	0.00	0.00	-0.01
24	3836	0.00	0.00	-0.01

Table 16: Level Difference Matrix for rising F1 of /i/

40 ms transition

		time (ms) from start of transition						
channel	frequency	0	5	10	15	20	25	30
1	111	0.00	-0.01	-0.01	-0.02	-0.03	-0.06	-0.08
2	152	0.00	-0.07	-0.20	-0.32	-0.52	-0.58	-0.74
3	196	0.00	0.00	-0.05	-0.14	-0.26	-0.40	-0.54
4	246	0.00	0.06	0.15	0.03	-0.05	-0.21	-0.20
5	300	0.00	0.05	0.21	0.47	0.75	1.11	1.44
6	360	0.01	0.28	0.42	0.91	1.12	1.56	1.78
7	427	0.04	0.18	0.30	0.62	0.79	1.13	1.28
8	500	0.06	0.15	0.28	0.52	0.68	0.92	1.05
9	581	0.05	0.09	0.26	0.45	0.58	0.78	0.96
10	670	0.03	0.06	0.17	0.37	0.50	0.71	0.91
11	769	0.03	0.05	0.27	0.40	0.51	0.61	0.81
12	878	0.02	0.03	0.17	0.26	0.36	0.40	0.61
13	1000	0.02	-0.01	0.14	0.21	0.32	0.35	0.55
14	1135	0.02	0.03	0.10	0.08	0.23	0.28	0.50
15	1286	0.13	0.17	0.17	0.18	0.31	0.31	0.37
16	1454	0.05	0.04	0.11	0.12	0.25	0.25	0.26
17	1642	0.10	0.11	0.04	0.03	0.14	0.15	0.21
18	1852	0.02	0.02	0.05	0.05	0.02	0.02	0.04
19	2089	0.01	0.01	0.01	0.01	0.01	0.01	-0.01
20	2356	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	2657	0.00	0.00	0.00	0.00	0.00	0.00	-0.01
22	2999	0.00	0.00	0.00	0.00	0.00	0.00	-0.01
23	3389	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24	3836	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01

Table 17: Level Difference Matrix for rising F1 of /i/

80 ms transition

		time (ms) from start of transition											
ch	freq	15	20	25	30	35	40	45	50	55	60	65	70
1	111	.00	.00	-.02	-.03	-.05	-.05	-.05	-.05	-.06	-.08	-.09	-.09
2	152	-.09	-.15	-.18	-.26	-.26	-.32	-.34	-.38	-.39	-.42	-.40	-.38
3	196	-.04	-.07	-.11	-.16	-.21	-.26	-.31	-.35	-.39	-.43	-.45	-.46
4	246	.02	.01	-.04	-.05	-.11	-.14	-.18	-.15	-.21	-.26	-.28	-.26
5	300	.12	.21	.33	.44	.57	.68	.81	.92	1.04	1.10	1.15	1.18
6	360	.24	.30	.50	.57	.69	.77	.92	.97	1.08	1.12	1.09	1.08
7	427	.18	.23	.36	.42	.51	.55	.67	.73	.80	.82	.83	.83
8	500	.15	.21	.32	.35	.38	.43	.55	.60	.65	.66	.66	.65
9	581	.11	.18	.27	.31	.32	.38	.45	.49	.51	.52	.55	.57
10	670	.07	.13	.25	.25	.25	.30	.38	.41	.46	.46	.48	.51
11	769	.04	.19	.27	.23	.26	.35	.40	.43	.45	.43	.40	.42
12	878	.08	.09	.07	.14	.20	.20	.23	.31	.33	.29	.28	.32
13	1000	.05	.09	.06	.12	.15	.23	.22	.19	.22	.25	.23	.25
14	1135	.00	.04	.06	.05	.05	.14	.17	.18	.18	.24	.26	.30
15	1286	.01	.13	.14	.11	.12	.24	.28	.29	.28	.25	.25	.28
16	1454	.03	.09	.10	.14	.14	.15	.13	.07	.08	.17	.18	.19
17	1642	.03	.10	.09	.06	.07	.14	.13	.10	.09	.10	.10	.19
18	1852	-.01	.02	.02	.01	.01	.03	.04	.02	.02	.00	.00	.02
19	2089	.00	.00	.01	.00	.00	.01	.01	.01	.01	.01	.00	-.01
20	2356	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
21	2657	.00	-.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
22	2999	.00	-.01	-.01	.00	.00	.00	.00	.00	.00	-.01	-.01	.00
23	3389	.00	-.01	-.01	-.01	-.01	-.01	-.01	.00	.00	.00	.00	-.01
24	3836	.00	-.01	-.01	-.01	-.01	.00	.00	.00	.00	.00	.00	.00

Table 18: Level Difference Matrix for rising F2 of /i/

20 ms transition

channel	frequency	time (ms) from start of transition 0	5	10
1	111	0.00	0.00	0.00
2	152	0.00	0.00	0.00
3	196	0.00	0.00	0.00
4	246	0.00	0.00	0.00
5	300	0.00	0.00	0.00
6	360	0.00	0.00	0.00
7	427	0.00	0.00	-0.01
8	500	0.00	-0.01	-0.02
9	581	-0.01	-0.03	-0.02
10	670	-0.01	-0.03	-0.04
11	769	-0.01	-0.04	-0.05
12	878	-0.08	-0.12	-0.14
13	1000	-0.11	-0.17	-0.25
14	1135	-0.07	-0.07	-0.27
15	1286	-0.06	-0.06	-0.30
16	1454	-0.24	-0.26	-0.67
17	1642	-0.44	-0.46	-1.21
18	1852	-1.05	-1.13	-2.60
19	2089	-1.32	-1.77	-3.78
20	2356	-0.15	-0.19	-0.50
21	2657	0.13	0.13	0.08
22	2999	-0.04	-0.03	-0.12
23	3389	-0.01	-0.01	-0.05
24	3836	0.00	0.01	0.03

Table 19: Level Difference Matrix for rising F2 of /i/

40 ms transition

		time (ms) from start of transition						
channel	frequency	0	5	10	15	20	25	30
1	111	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	152	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	196	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	246	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	300	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	360	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	427	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00
8	500	0.00	0.00	0.00	0.00	0.00	0.00	-0.01
9	581	0.00	0.00	-0.01	-0.02	-0.03	-0.04	-0.02
10	670	-0.02	-0.07	-0.06	-0.06	-0.05	-0.04	-0.05
11	769	0.00	-0.02	-0.02	0.00	-0.01	-0.02	-0.05
12	878	-0.05	-0.09	-0.12	-0.10	-0.12	-0.15	-0.15
13	1000	-0.03	-0.05	-0.16	-0.21	-0.24	-0.24	-0.24
14	1135	-0.09	-0.10	-0.15	-0.18	-0.19	-0.22	-0.35
15	1286	-0.05	-0.06	-0.10	-0.10	-0.19	-0.17	-0.26
16	1454	-0.07	-0.08	-0.29	-0.31	-0.44	-0.46	-0.69
17	1642	-0.13	-0.14	-0.55	-0.56	-0.94	-0.96	-1.28
18	1852	-0.42	-0.46	-1.21	-1.23	-2.02	-2.05	-2.54
19	2089	-0.54	-0.73	-1.83	-2.04	-3.05	-3.18	-3.70
20	2356	-0.05	-0.06	-0.34	-0.42	-0.37	-0.43	-0.44
21	2657	0.05	0.05	0.13	0.11	0.25	0.23	0.02
22	2999	-0.02	-0.02	-0.05	-0.04	-0.07	-0.06	-0.11
23	3389	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.05
24	3836	0.00	0.00	0.01	0.01	0.02	0.02	0.02

Table 20: Level Difference Matrix for rising F2 of /ɪ/

80 ms transition

time (ms) from start of transition

ch	freq	15	20	25	30	35	40	45	50	55	60	65	70
1	111	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
2	152	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
3	196	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
4	246	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
5	300	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
6	360	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
7	427	.00	.00	.00	.00	.00	.00	-.01	-.01	.00	.00	.00	.00
8	500	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	-.01
9	581	.00	.00	.00	.00	.00	.00	.00	-.01	-.02	-.03	-.04	-.02
10	670	.00	.00	.00	.00	.01	-.02	-.07	-.06	-.06	-.05	-.04	-.05
11	769	.00	.00	.00	.00	-.01	.00	-.02	-.02	.00	-.01	-.02	-.05
12	878	.00	.00	.00	.00	.00	-.05	-.09	-.12	-.10	-.12	-.15	-.15
13	1000	.00	.00	.00	.00	.00	-.03	-.05	-.16	-.21	-.24	-.24	-.24
14	1135	.00	.00	.00	.00	.00	-.09	-.10	-.15	-.18	-.19	-.22	-.35
15	1286	.00	.00	.00	.00	.00	-.05	-.06	-.10	-.10	-.19	-.17	-.26
16	1454	.00	.00	.00	.00	.00	-.07	-.08	-.29	-.31	-.44	-.46	-.69
17	1642	.00	.00	.00	.00	.00	-.13	-.14	-.55	-.56	-.94	-.96	-1.28
18	1852	.00	.00	.00	.00	.00	-.42	-.46	-1.21	-1.23	-2.02	-2.05	-2.54
19	2089	.00	.00	.00	.00	.00	-.54	-.73	-1.83	-2.04	-3.05	-3.18	-3.70
20	2356	.00	.00	.00	.00	.00	-.05	-.06	-.34	-.42	-.37	-.43	-.44
21	2657	.00	.00	.00	.00	.00	.05	.05	.13	.11	.25	.23	.02
22	2999	.00	.00	.00	.00	.00	-.02	-.02	-.05	-.04	-.07	-.06	-.11
23	3389	.00	.00	.00	.00	.00	-.01	-.01	-.01	-.01	-.02	-.02	-.05
24	3836	.00	.00	.00	.00	.00	.00	.00	.01	.01	.02	.02	.02

Table 21: Level Difference Matrix for rising isolated F2 of /i/

20 ms transition

		time (ms) from start of transition		
channel	frequency	0	5	10
1	111	0.00	0.00	0.00
2	152	0.00	0.00	0.00
3	196	0.00	0.00	0.00
4	246	0.00	0.00	0.00
5	300	0.00	0.00	0.00
6	360	0.00	0.00	-0.01
7	427	0.00	0.00	0.00
8	500	0.00	0.00	-0.01
9	581	0.00	0.00	-0.01
10	670	0.00	0.00	-0.01
11	769	-0.01	-0.02	-0.03
12	878	-0.02	-0.03	-0.05
13	1000	-0.03	-0.05	-0.08
14	1135	-0.04	-0.05	-0.09
15	1286	-0.06	-0.07	-0.14
16	1454	-0.09	-0.09	-0.22
17	1642	-0.14	-0.14	-0.38
18	1852	-0.32	-0.34	-0.83
19	2089	-0.63	-0.84	-1.64
20	2356	0.03	0.03	-0.41
21	2657	0.64	0.73	1.41
22	2999	0.45	0.47	1.19
23	3389	0.30	0.30	0.82
24	3836	0.24	0.24	0.65

Table 22: Level Difference Matrix for rising isolated F2 of /i/

40 ms transition

		time (ms) from start of transition						
channel	frequency	0	5	10	15	20	25	30
1	111	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	152	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	196	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	246	0.00	0.00	0.01	0.00	0.00	0.00	0.00
5	300	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	360	0.00	0.00	0.00	0.00	0.00	-0.01	0.00
7	427	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	500	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
9	581	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01
10	670	0.00	0.00	-0.01	-0.01	-0.01	-0.02	-0.02
11	769	-0.01	-0.01	-0.01	-0.01	-0.02	-0.03	-0.03
12	878	-0.01	-0.01	-0.02	-0.03	-0.03	-0.03	-0.04
13	1000	-0.01	-0.02	-0.03	-0.03	-0.05	-0.05	-0.07
14	1135	-0.02	-0.02	-0.03	-0.04	-0.06	-0.06	-0.07
15	1286	-0.02	-0.03	-0.06	-0.06	-0.08	-0.08	-0.11
16	1454	-0.03	-0.03	-0.07	-0.08	-0.11	-0.11	-0.17
17	1642	-0.04	-0.04	-0.12	-0.12	-0.19	-0.19	-0.27
18	1852	-0.09	-0.09	-0.27	-0.28	-0.43	-0.44	-0.60
19	2089	-0.18	-0.24	-0.59	-0.67	-1.04	-1.11	-1.26
20	2356	0.02	0.03	-0.13	-0.18	-0.34	-0.39	-0.56
21	2657	0.18	0.21	0.41	0.42	0.67	0.70	0.88
22	2999	0.13	0.13	0.32	0.32	0.52	0.53	0.77
23	3389	0.09	0.09	0.24	0.24	0.37	0.37	0.55
24	3836	0.09	0.09	0.17	0.17	0.31	0.31	0.43

Table 23: Level Difference Matrix for rising isolated F2 of /i/

80 ms transition

ch	freq	time (ms) from start of transition											
		15	20	25	30	35	40	45	50	55	60	65	70
1	111	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
2	152	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
3	196	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
4	246	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
5	300	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
6	360	.00	.00	.00	.00	.00	.00	-.01	-.01	.00	.00	.00	.00
7	427	.00	.00	.00	.00	-.01	-.01	-.01	-.01	.00	.00	.00	.00
8	500	-.01	-.01	-.01	-.01	-.01	-.01	-.01	-.01	-.01	-.02	-.02	-.02
9	581	.00	.00	.00	.00	.00	-.01	-.01	-.01	.00	-.01	-.01	-.01
10	670	.00	-.01	-.01	-.01	-.01	-.01	-.01	-.02	-.02	-.02	-.02	-.02
11	769	-.01	-.01	-.01	-.02	-.02	-.03	-.02	-.03	-.04	-.04	-.03	-.03
12	878	-.01	-.02	-.02	-.03	-.04	-.03	-.03	-.04	-.04	-.05	-.05	-.05
13	1000	-.02	-.03	-.03	-.04	-.04	-.04	-.04	-.06	-.06	-.06	-.06	-.06
14	1135	-.02	-.04	-.04	-.05	-.05	-.06	-.06	-.07	-.07	-.08	-.07	-.07
15	1286	-.04	-.06	-.06	-.08	-.08	-.09	-.09	-.10	-.10	-.10	-.10	-.10
16	1454	-.04	-.08	-.08	-.10	-.10	-.13	-.14	-.14	-.14	-.15	-.15	-.15
17	1642	-.08	-.11	-.11	-.17	-.17	-.21	-.21	-.23	-.23	-.24	-.24	-.25
18	1852	-.16	-.27	-.27	-.38	-.38	-.49	-.49	-.54	-.54	-.55	-.55	-.54
19	2089	-.41	-.64	-.69	-.93	-.97	-1.21	-1.25	-1.37	-1.37	-1.38	-1.38	-1.20
20	2356	-.10	-.22	-.25	-.35	-.37	-.43	-.45	-.48	-.48	-.48	-.48	-.60
21	2657	.26	.39	.40	.56	.57	.77	.79	.89	.89	.90	.90	.80
22	2999	.19	.31	.31	.44	.44	.59	.60	.68	.68	.68	.68	.69
23	3389	.14	.23	.23	.32	.32	.43	.43	.48	.48	.47	.47	.49
24	3836	.12	.20	.20	.26	.26	.35	.35	.37	.37	.38	.38	.38

Table 24: Level Difference Matrix for rising F1 of /a/

20 ms transition

		time (ms) from start of transition		
channel	frequency	0	5	10
1	111	0.00	0.00	-0.01
2	152	0.00	-0.05	-0.07
3	196	-0.01	-0.04	-0.07
4	246	0.01	-0.06	-0.10
5	300	-0.02	-0.11	-0.18
6	360	-0.05	-0.26	-0.41
7	427	-0.05	-0.29	-0.40
8	500	-0.13	-0.45	-0.60
9	581	-0.20	-0.85	-1.07
10	670	-0.05	-1.61	-1.73
11	769	-0.09	-0.68	-1.18
12	878	-0.05	0.79	1.05
13	1000	0.01	0.26	0.24
14	1135	-0.01	-0.02	-0.19
15	1286	-0.09	-0.13	-0.52
16	1454	-0.20	-0.23	-0.81
17	1642	-0.27	-0.28	-0.94
18	1852	-0.27	-0.27	-0.78
19	2089	-0.11	-0.11	-0.32
20	2356	-0.01	-0.01	-0.05
21	2657	0.01	0.01	0.04
22	2999	0.02	0.03	0.10
23	3389	0.01	0.01	0.03
24	3836	-0.01	-0.01	-0.04

Table 25: Level Difference Matrix for rising F1 of /a/

40 ms transition

		time (ms) from start of transition						
channel	frequency	0	5	10	15	20	25	30
1	111	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01
2	152	0.00	-0.02	-0.02	-0.03	-0.04	-0.07	-0.09
3	196	0.00	-0.01	-0.02	-0.04	-0.06	-0.08	-0.11
4	246	0.01	-0.03	-0.04	-0.05	-0.08	-0.13	-0.15
5	300	-0.01	-0.05	-0.07	-0.12	-0.15	-0.22	-0.25
6	360	0.00	-0.08	-0.13	-0.24	-0.28	-0.40	-0.46
7	427	-0.02	-0.08	-0.12	-0.25	-0.31	-0.44	-0.48
8	500	-0.04	-0.13	-0.21	-0.40	-0.51	-0.70	-0.71
9	581	-0.06	-0.25	-0.36	-0.68	-0.82	-1.15	-1.19
10	670	-0.01	-0.47	-0.50	-1.08	-1.11	-1.73	-1.78
11	769	-0.03	-0.16	-0.22	-0.29	-0.37	-0.55	-0.78
12	878	-0.01	0.24	0.21	0.45	0.44	0.50	0.75
13	1000	0.01	0.07	0.06	0.18	0.17	0.29	0.32
14	1135	0.00	-0.01	-0.02	-0.03	-0.06	-0.06	-0.14
15	1286	-0.02	-0.03	-0.10	-0.11	-0.17	-0.19	-0.36
16	1454	-0.05	-0.06	-0.17	-0.19	-0.32	-0.34	-0.53
17	1642	-0.08	-0.08	-0.24	-0.25	-0.40	-0.40	-0.61
18	1852	-0.08	-0.08	-0.24	-0.24	-0.33	-0.33	-0.52
19	2089	-0.05	-0.05	-0.07	-0.08	-0.14	-0.14	-0.23
20	2356	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.04
21	2657	0.00	0.00	0.01	0.01	0.01	0.01	0.03
22	2999	0.01	0.01	0.03	0.03	0.05	0.06	0.06
23	3389	0.00	0.01	0.01	0.01	0.02	0.02	0.03
24	3836	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.02

Table 26: Level Difference Matrix for rising F1 of /a/

80 ms transition

time (ms) from start of transition

ch	freq	15	20	25	30	35	40	45	50	55	60	65	70
1	111	.01	.01	.01	.01	.01	.01	.00	.00	.00	.00	.00	.00
2	152	.00	-.02	-.02	-.02	-.03	-.05	-.05	-.06	-.04	-.04	-.06	-.06
3	196	.00	-.01	-.02	-.03	-.04	-.05	-.06	-.07	-.09	-.09	-.09	-.09
4	246	.00	-.04	-.05	-.05	-.06	-.07	-.10	-.11	-.11	-.12	-.12	-.12
5	300	-.04	-.04	-.06	-.09	-.12	-.14	-.18	-.20	-.21	-.22	-.22	-.22
6	360	-.09	-.10	-.15	-.17	-.22	-.25	-.31	-.33	-.34	-.35	-.36	-.36
7	427	-.13	-.15	-.18	-.19	-.24	-.26	-.33	-.36	-.41	-.42	-.41	-.39
8	500	-.16	-.22	-.31	-.35	-.43	-.48	-.55	-.57	-.63	-.65	-.65	-.59
9	581	-.28	-.33	-.49	-.55	-.70	-.76	-.90	-.95	-1.04	-1.05	-1.04	-.98
10	670	-.45	-.46	-.71	-.72	-1.00	-1.00	-1.29	-1.28	-1.38	-1.37	-1.38	-1.43
11	769	-.08	-.09	-.09	-.11	-.15	-.19	-.28	-.34	-.39	-.40	-.41	-.46
12	878	.23	.22	.32	.31	.37	.37	.40	.41	.33	.33	.30	.45
13	1000	.09	.09	.14	.14	.19	.19	.24	.24	.26	.26	.26	.28
14	1135	.00	-.01	-.01	-.02	-.02	-.03	-.03	-.04	-.05	-.05	-.05	-.09
15	1286	-.04	-.05	-.06	-.09	-.09	-.13	-.14	-.16	-.16	-.16	-.16	-.22
16	1454	-.07	-.10	-.11	-.18	-.19	-.24	-.25	-.27	-.28	-.29	-.29	-.32
17	1642	-.09	-.14	-.14	-.24	-.25	-.28	-.28	-.36	-.37	-.34	-.34	-.38
18	1852	-.05	-.12	-.12	-.22	-.22	-.24	-.24	-.33	-.34	-.31	-.31	-.29
19	2089	-.02	-.05	-.05	-.07	-.07	-.11	-.11	-.13	-.13	-.11	-.11	-.11
20	2356	.01	.00	.00	.00	.01	.00	.00	.00	.00	.01	.00	-.01
21	2657	.02	.02	.02	.02	.02	.03	.03	.03	.03	.03	.03	.02
22	2999	.02	.04	.04	.04	.04	.05	.05	.06	.06	.04	.05	.06
23	3389	.02	.02	.02	.03	.03	.02	.02	.03	.03	.03	.03	.03
24	3836	.01	.01	.01	.01	.00	-.01	-.01	.00	.00	.00	.00	.01

Table 27: Level Difference Matrix for rising F2 of /a/

20 ms transition

channel	frequency	time (ms) from start of transition 0	5	10
1	111	-0.65	-0.65	-0.65
2	152	-0.66	-0.66	-0.65
3	196	-0.66	-0.66	-0.65
4	246	-0.66	-0.66	-0.65
5	300	-0.67	-0.65	-0.63
6	360	-0.64	-0.60	-0.57
7	427	-0.64	-0.59	-0.57
8	500	-0.63	-0.59	-0.56
9	581	-0.64	-0.59	-0.55
10	670	-0.65	-0.61	-0.59
11	769	-0.66	-0.66	-0.65
12	878	-0.65	-0.76	-0.76
13	1000	-0.62	-0.86	-1.07
14	1135	-0.47	0.09	0.68
15	1286	0.00	0.32	1.22
16	1454	-0.05	0.02	1.11
17	1642	-0.20	-0.18	0.60
18	1852	-0.36	-0.35	0.19
19	2089	-0.51	-0.51	-0.28
20	2356	-0.62	-0.63	-0.61
21	2657	-0.66	-0.66	-0.70
22	2999	-0.69	-0.69	-0.80
23	3389	-0.67	-0.67	-0.70
24	3836	-0.66	-0.66	-0.64

Table 28: Level Difference Matrix for rising F2 of /a/

40 ms transition

		time (ms) from start of transition						
channel	frequency	0	5	10	15	20	25	30
1	111	-0.72	-0.73	-0.73	-0.73	-0.73	-0.73	-0.74
2	152	-0.74	-0.75	-0.75	-0.72	-0.72	-0.73	-0.74
3	196	-0.74	-0.74	-0.74	-0.74	-0.74	-0.72	-0.72
4	246	-0.74	-0.74	-0.74	-0.73	-0.73	-0.70	-0.70
5	300	-0.74	-0.73	-0.73	-0.71	-0.71	-0.69	-0.68
6	360	-0.73	-0.72	-0.71	-0.67	-0.67	-0.63	-0.60
7	427	-0.74	-0.71	-0.69	-0.67	-0.65	-0.61	-0.59
8	500	-0.73	-0.71	-0.70	-0.67	-0.65	-0.61	-0.59
9	581	-0.72	-0.71	-0.70	-0.66	-0.64	-0.60	-0.57
10	670	-0.73	-0.72	-0.71	-0.68	-0.67	-0.64	-0.62
11	769	-0.74	-0.75	-0.74	-0.74	-0.74	-0.74	-0.74
12	878	-0.74	-0.79	-0.79	-0.86	-0.86	-0.92	-0.94
13	1000	-0.72	-0.83	-0.89	-1.01	-1.05	-1.17	-1.31
14	1135	-0.66	-0.41	-0.10	0.31	0.64	1.03	1.36
15	1286	-0.46	-0.32	0.21	0.42	0.93	1.17	1.49
16	1454	-0.49	-0.46	0.07	0.11	0.62	0.64	1.14
17	1642	-0.57	-0.57	-0.19	-0.18	0.21	0.22	0.59
18	1852	-0.62	-0.62	-0.42	-0.42	-0.14	-0.13	0.12
19	2089	-0.68	-0.68	-0.58	-0.58	-0.50	-0.50	-0.38
20	2356	-0.73	-0.73	-0.73	-0.73	-0.70	-0.70	-0.71
21	2657	-0.74	-0.74	-0.75	-0.75	-0.76	-0.76	-0.80
22	2999	-0.76	-0.76	-0.77	-0.77	-0.82	-0.82	-0.87
23	3389	-0.73	-0.73	-0.73	-0.73	-0.75	-0.75	-0.76
24	3836	-0.75	-0.75	-0.73	-0.73	-0.74	-0.74	-0.71

Table 29: Level Difference Matrix for rising F2 of /a/

80 ms transition

time (ms) from start of transition

ch	freq	15	20	25	30	35	40	45	50	55	60	65	70
1	111	-.55	-.55	-.55	-.55	-.54	-.54	-.54	-.54	-.55	-.55	-.55	-.55
2	152	-.55	-.55	-.55	-.54	-.54	-.54	-.54	-.53	-.54	-.55	-.54	-.54
3	196	-.56	-.55	-.55	-.54	-.54	-.54	-.53	-.53	-.52	-.53	-.53	-.53
4	246	-.55	-.55	-.54	-.53	-.54	-.54	-.53	-.51	-.50	-.51	-.52	-.52
5	300	-.54	-.54	-.54	-.53	-.52	-.51	-.50	-.50	-.49	-.49	-.49	-.49
6	360	-.53	-.52	-.50	-.49	-.47	-.47	-.46	-.46	-.43	-.43	-.45	-.45
7	427	-.54	-.52	-.49	-.48	-.48	-.47	-.46	-.45	-.43	-.43	-.43	-.43
8	500	-.52	-.52	-.50	-.48	-.46	-.45	-.43	-.42	-.42	-.42	-.41	-.41
9	581	-.52	-.51	-.49	-.48	-.45	-.44	-.43	-.42	-.41	-.41	-.41	-.40
10	670	-.54	-.53	-.51	-.50	-.48	-.48	-.47	-.47	-.46	-.46	-.46	-.46
11	769	-.56	-.56	-.56	-.56	-.56	-.56	-.56	-.56	-.56	-.56	-.56	-.56
12	878	-.62	-.62	-.65	-.65	-.69	-.69	-.72	-.72	-.73	-.73	-.73	-.75
13	1000	-.68	-.72	-.77	-.80	-.85	-.87	-.92	-.93	-.92	-.92	-.92	-1.03
14	1135	-.05	.11	.33	.50	.71	.87	1.06	1.19	1.29	1.32	1.33	1.40
15	1286	.00	.27	.37	.63	.74	.98	1.10	1.22	1.27	1.28	1.28	1.26
16	1454	-.18	.10	.13	.38	.40	.67	.68	.81	.81	.81	.81	.82
17	1642	-.31	-.11	-.11	.11	.11	.32	.33	.40	.41	.40	.40	.40
18	1852	-.39	-.28	-.28	-.14	-.14	.01	.01	.06	.06	.04	.05	.06
19	2089	-.47	-.45	-.45	-.39	-.39	-.35	-.35	-.31	-.31	-.30	-.30	-.30
20	2356	-.55	-.55	-.55	-.54	-.53	-.53	-.53	-.53	-.53	-.52	-.52	-.53
21	2657	-.57	-.57	-.57	-.57	-.57	-.58	-.58	-.58	-.58	-.58	-.58	-.59
22	2999	-.59	-.60	-.60	-.62	-.62	-.61	-.61	-.62	-.62	-.64	-.63	-.66
23	3389	-.56	-.57	-.57	-.56	-.56	-.56	-.56	-.56	-.56	-.57	-.57	-.57
24	3836	-.56	-.56	-.56	-.55	-.55	-.56	-.56	-.54	-.54	-.56	-.56	-.53

REFERENCES

Ainsworth, W.A. (1968). First formant transitions and the perception of synthetic semivowels. *JASA.* **44**, 689-694.

Ainsworth, W.A. (1974). Performance of a speech synthesis system. *Int. J. Man-Machine Studies.* **6**, 493-511.

Ainsworth, W.A. (1976). *Mechanisms of Speech Recognition.* Pergamon Press, Oxford.

Bischof, W.F., Foster, D.H., and Kahn, J.I. (1985). Selective internal operations in the recognition of locally and globally point-inverted patterns. *Spatial Vision.* **1**, 179-196.

Brady, P.T., House, A.S., and Stevens, K.N. (1961). Perception of sounds characterized by a rapidly changing resonant frequency. *JASA.* **33**, 1357-1362.

Carlyon, R.P., and Moore, B.C.J. (1984). Intensity discrimination: A severe departure from Weber's law. *JASA.* **76**, 1369-1376.

Chistovich, L.A., Lublinskaya, V.V., Malinnikova, T.G., Ogorodnikova, E.A., Stoljarova, E.I., and Zhukov, S.J. (1982). Temporal processing of peripheral auditory patterns of speech. In *The Representation of Speech in the Peripheral Auditory System,* (eds R. Carlson and B. Granström). Elsevier Biomedical Press, Amsterdam.

Collins, M.J., and Cullen, J.K., Jr. (1978). Temporal integration of tone glides. *JASA.* **63**, 469-473.

Collins, M.J., and Cullen, J.K., Jr. (1984). Effects of background noise level on detection of tone glides. *JASA.* **76**, 1696-1698.

Cosgrove, P., and Wilson, J.P. (1986). A study of frequency transition detection using synthetic vowel sounds. *Proc. IOA Autumn Conf.* **8**, 9-16.

Cullen, J .K., Jr., and Collins, M.J. (1982). Audibility of short-duration tone-glides as a function of rate of frequency change. *Hearing Research.* **7**, 115-125.

Danaher, E.M., Osberger, M.J., and Pickett, J.M. (1973). Discrimination of formant frequency transitions in synthetic vowels. *J. Speech Hear. Res.* **16**, 439-451.

Darwin, C.J. (1971). Dichotic backward masking of complex sounds. *Q. J. Exptl. Psychol.* **23**, 386-392.

Delattre, P.C., Liberman, A.M., and Cooper, F.S. (1955). Acoustic loci and transitional cues for consonants. *JASA.* **27**, 769-773.

Demany, L. (1983). Speeded discrimination of frequency and intensity differences. *Proc. 11th Int. Cong. on Acoustics.* **3**, 107-110.

Diehl, R.L. (1975). The effect of selective adaptation on the identification of speech sounds. *Percept. Psychophys.* **17**, 48-52.

Dunn, H.K. (1961). Methods of measuring vowel formant bandwidths. *JASA.* **33**, 1737-1746.

Egan, J.P., and Hake, H.W. (1950). On the masking pattern of a simple auditory stimulus. *JASA.* **22**, 622-630.

Elliot, L.L. (1971). Backward and forward masking. *Audiology.* **10**, 65-76.

Evans, E.F. (1972). The frequency response and other properties of single fibres in the guinea-pig cochlear nerve. *J. Physiol.* **226**, 263-287.

Evans, E.F. (1981). Neural encoding of speech signals at peripheral and central levels of the auditory system. In *The Cognitive Representation of Speech*, (eds T. Myers, J. Laver and J. Anderson). North-Holland Publishing Co.

Evans, E.F. (1982). Functions of the auditory system. In *The Senses*, (eds H.B. Barlow and J.D. Mollon). Cambridge University Press.

Evans, E.F., and Palmer, A.R. (1975). Responses of units in the cochlear nerve and nucleus of the cat to signals in the presence of bandstop noise. *J. Physiol.* **252**, 60-62P.

Evans, E.F., and Palmer, A.R. (1980). Relationship between dynamic range of cochlear nerve fibres and their spontaneous activity. *Exptl. Brain Res.* **40**, 115-118.

Fastl, H. (1983). Fluctuation strength of FM-tones. *Proc. 11th Int. Cong. on Acoustics.* **3**, 123-126.

Flanagan, J.L. (1955a). A difference limen for vowel formant frequency. *JASA.* **27**, 613-617 [L].

Flanagan, J.L. (1955b). Difference limen for the intensity of a vowel sound. *JASA.* **27**, 1223-1225 [L].

Flanagan, J.L. (1957). Difference limen for formant amplitude. *J. Speech Hearing Disorders.* **22**, 205-212.

Fletcher, H. (1940). Auditory patterns. *Rev. Mod. Phys.* **12**, 47-65.

Foster, D.H. (1986). Estimating the variance of a critical stimulus level from sensory performance data. *Biol. Cybern.* **53**, 189-194.

Fry, D.B. (1955). Duration and intensity as physical correlates of linguistic stress. *JASA.* **27**, 765-768.

Fry, D.B. (1958). Experiments in the perception of stress. *Language and Speech.* **1**, 126-152.

Fujisaki, H., and Sekimoto, (1975). Perception of time-varying resonances in speech and non-speech stimuli. In *Structure and Process in Speech Perception*, (eds A. Cohen and S.G. Nooteboom). Springer-Verlag, Berlin.

Gardner, R.B., and Wilson, J.P. (1979). Evidence for direction-specific channels in the processing of frequency modulation. *JASA.* **66**, 704-709.

Gay, T. (1970). A perceptual study of American English diphthongs. *Language and Speech.* **13**, 65-88.

Goldstein, J.L., and Srulovicz, P. (1977). Auditory-nerve spike intervals as an adequate basis for aural frequency measurement. In *Psychophysics and Physiology of Hearing*, (eds E.F. Evans and J.P. Wilson). Academic Press, London.

Haggard, M. (1977). Mechanisms of formant frequency discrimination. In *Psychophysics and Physiology of Hearing*, (eds E.F. Evans and J.P. Wilson). Academic Press, London.

Harris, J.D. (1952). Pitch discrimination. *JASA.* **24**, 750-755.

Harris, J.D., Pikler, A.G., Hoffman, H.S., and Ehmer, R.H. (1958). The interaction of pitch and loudness discriminations. *J. Exptl. Psychol.* **56**, 232-238.

Henning, G.B. (1966). Frequency discrimination of random amplitude tones. *JASA.* **39**, 336-339.

Hoekstra, A. (1979). *Frequency Discrimination and Frequency Analysis in Hearing.* PhD thesis, University of Groningen.

Holmes, J.N. (1982). Formant synthesizers: cascade or parallel? *JSRU Research Report* No. 1017.

Holmes, J.N., Mattingly, I.G., and Shearme, J.N. (1964). Speech synthesis by rule. *Language and Speech.* **7**, 127-143.

Holmgren, K. (1979). Formant frequency target versus rate of change in vowel identification. *Phonet. Exper. Res. Inst. Ling. Univ. Stockholm.* **1**, 83-91.

Horst, J.W. (1982). *Discrimination of Complex Signals in Hearing.* PhD thesis, University of Groningen.

House, A.S. (1960). Formant bandwidths and vowel preference. *J. Speech Hear. Res.* **3**, 3-8.

Houtgast, T. (1972). Psychophysical evidence for lateral inhibition in hearing. *JASA.* **51**, 1885-1894.

Hughes, J.W. (1946). The threshold of audition for short periods of stimulation. *Proc. R. Soc.* **B133**, 486-490.

Jeffress, L.A. (1970). Masking. In *Foundations of Modern Auditory Theory*, Vol. 1 (ed. J.V. Tobias). Academic Press, New York.

Jesteadt, W., Wier, C.C., and Green, D.M. (1977). Intensity discrimination as a function of frequency and sensation level. *JASA.* **61**, 169-177.

Kay, R.H., and Matthews, D.R. (1972). On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones. *J. Physiol.* **225**, 657-677.

Kay, R.N. (1987). The Alvey Directorate: The U.K. Alvey Research Programme in Speech Technology. *Acoustics Bulletin.* Vol. **12**, No. 1, 4-8.

Kiang, N.Y.S. (1968). A survey of recent developments in the study of auditory physiology. *Ann. Otol. Rhinol. Laryngol.* **77**, 656-675.

Lacerda, F., and Moreira, O. (1982). How does the peripheral auditory system represent formant transitions? In *The Representation of Speech in the Peripheral Auditory System*, (eds R. Carlson and B. Granström). Elsevier Biomedical Press, Amsterdam.

Lehiste, I., and Peterson, G.E. (1961). Transitions, glides and diphthongs. *JASA.* **33**, 268-277.

Liberman, A.M. (1984). Brief comments on invariance in phonetic perception. *Status Report on Speech Research.* Haskins Labs, Conn. USA SR-77/78, 27-30.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* **74**, 431-461.

Liberman, A.M., Delattre, P.C., and Cooper, F.S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech.* 1, 153-167.

Liberman, A.M., Delattre, P.C., Cooper, F.S., and Gerstman, L.J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychol. Monogr.* **68**, No. 8, (Whole No. 379).

Liberman, A.M., Delattre, P.C., Gerstman, L.J., and Cooper, F.S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *J. Exptl. Psychol.* **52**, 127-137.

Licklider, J.C.R. (1951). Basic correlates of the auditory stimulus. In *Handbook of Experimental Psychology*, (ed. S.S. Stevens). Wiley, New York.

Lindblom, B.E.F., and Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *JASA.* **42**, 830-843.

Linggard, R. (1985). *Electronic Synthesis of Speech.* Cambridge University Press, Cambridge.

Lynn, P.A. (1982). *An Introduction to the Analysis and Processing of Signals.* 2nd Edn. Macmillan, London.

McClelland, K.D., and Brandt, J.F. (1969). Pitch of frequency-modulated sinusoids. *JASA.* **45**, 1489-1498.

Marslen-Wilson, W.D. (1984). Function and process in spoken word-recognition. In *Attention and Performance X: Control of Language Processes*, (eds H. Bouma and D.G. Bouwhuis). Erlbaum, Hillsdale New Jersey.

Mattingly, I.G., and Liberman, A.M. (1986). Specialized perceiving systems for speech and other biologically significant sounds. *Status Report on Speech Research.* Haskins Labs, Conn. USA, SR86/87, 25-43.

Mattingly, I.G., Liberman, A.M., Syrdal, A.K., and Halwes, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology.* **2**, 131-157.

Meddis, R. (1986). Exploiting physiological codes in ASR. *Proc. IOA Autumn Conf.* **8**, 581-588.

Mermelstein, P. (1978). Difference limens for formant frequencies of steady-state and consonant-bound vowels. *JASA.* **63**, 572-580.

Miller, G.A. (1947). Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *JASA.* **19**, 609-619.

Moore, B.C.J. (1973a). Frequency difference limens for short-duration tones. *JASA.* **54**, 610-619.

Moore, B.C.J. (1973b). Frequency difference limens for narrow bands of noise. *JASA.* **54**, 888-896.

Moore, B.C.J. (1974). Relation between the critical bandwidth and the frequency difference limen. *JASA.* **55**, 359 [L].

Moore, B.C.J. (1978). Psychophysical tuning curves measured in simultaneous and forward masking. *JASA.* **63**, 524-532.

Moore, B.C.J. (1982). *An Introduction to the Psychology of Hearing.* 2nd Edn. Academic Press, London.

Moore, B.C.J. (1983). Review paper: Psychoacoustics of normal and impaired listeners. In *Hearing- Physiological Bases and Psychophysics,* (eds R. Klinke and R. Hartmann). Springer-Verlag, Berlin.

Moore, B.C.J., and Glasberg, B.R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *JASA.* **74**, 750-753.

Moore, B.C.J., and Glasberg, B.R. (1986a). The role of frequency selectivity in the perception of loudness, pitch and time. In *Frequency Selectivity in Hearing,* (ed. B.C.J. Moore). Academic Press, London.

Moore, B.C.J., and Glasberg, B.R. (1986b). The relationship between frequency selectivity and frequency discrimination for subjects with unilateral and bilateral cochlear impairments. In *Auditory Frequency Selectivity,* (eds. B.C.J. Moore and R.D. Patterson). Plenum Press, New York.

Moore, B.C.J., and Raab, D.H. (1974). Pure-tone intensity discrimination: some experiments relating to the 'near-miss' to Weber's law. *JASA.* **55**, 1049-1054.

Moore, B.C.J., and Raab, D.H. (1975). Intensity discrimination for noise bursts in the presence of a continuous, bandstop background: effects of level, width of the bandstop, and duration. *JASA.* 57, 400-405.

Moore, R.K., Beardsley, D., Russell, M.J., and Tomlinson, M.J. (1982). Towards an integrated discriminative network for automatic speech recognition. *Proc. IOA Autumn Conf.* C1.1-1.5.

Moore, R.K., and Bridle, J.S. (1986). Speech research at RSRE. *Proc. IOA Autumn Conf.* 8, 257-264.

Moore, R.K., Russell, M.J., and Tomlinson, M.J. (1982). Locally constrained dynamic programming in automatic speech recognition. *Proc. IEEE Int. Conf. ASSP.* 1270-1272.

Nabelek, I., and Hirsh, I.J. (1969). On the discrimination of frequency transitions. *JASA.* 45, 1510-1519.

O'Connor, J.D., Gerstman, L.J., Liberman, A.M., Delattre, P.C., and Cooper, F.S. (1957). Acoustic cues for the perception of initial /w, j, r, l/ in English. *Word.* 13, 24-43.

Palmer, A.R., and Evans, E.F. (1979). On the peripheral coding of the level of individual frequency components of complex sounds at high sound levels. In *Hearing Mechanisms and Speech,* (eds O. Creutzfeldt, H. Scheich and C. Schreiner). Springer-Verlag, Berlin.

Palmer, A.R., and Russell, I.R. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing Research.* 24, 1-15.

Palmer, A.R., Winter, I.M., and Darwin, C.J. (1986). The representation of steady-state vowels in the temporal discharge patterns of guinea-pig cochlear nerve and primary like cochlear nucleus neurones. *JASA.* 79, 100-113.

Patterson, R.D. (1974). Auditory filter shape. *JASA.* 55, 802-809.

Patterson, R.D. (1976). Auditory filter shapes derived with noise stimuli. *JASA.* 59, 640-654.

Patterson, R.D. (1986). Spiral detection of periodicity and the spiral form of musical scales. *Psychology of Music.* 14, 44-61.

Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1987). An efficient auditory filterbank based on the gammatone function. Paper presented at a meeting of the IOA Speech Group on *Auditory Modelling* at RSRE, Malvern, December 14-15.

Peckham, J. (1986). When machines have ears. *New Scientist.* 4th December, No. 1537. 54-56.

Pick, G.F. (1980). Level dependence of psychophysical frequency resolution and auditory filter shape. *JASA.* 1085-1095.

Pick, G.F., Evans, E.F., and Wilson, J.P. (1977). Frequency resolution in patients with hearing loss of cochlear origin. In *Psychophysics and Physiology of Hearing,* (eds E.F. Evans and J.P. Wilson). Academic Press, London.

Pisoni, D.B. (1983). Speech perception: some new directions in research and theory. In *Research on Speech Perception.* Progress Report No. 9, Indiana University.

Plomp, R. (1964). Rate of decay of auditory sensation. *JASA.* **36**, 277-282.

Pols, L.C.W., Boxelaar, G.W., and Koopmans-Van Beinum, F.J. (1984). Study of the role of formant transitions in vowel recognition using the matching paradigm. *Proc. IOA Autumn Conf.* **6**, 371-378.

Pols, L.C.W., and Schouten, M.E.H. (1982). Perceptual relevance of coarticulation. In *The Representation of Speech in the Peripheral Auditory System,* (eds R. Carlson and B. Granström). Elsevier Biomedical Press, Amsterdam.

Repp, B.H. (1986). The role of psychophysics in understanding speech perception. *Status Report on Speech Research.* Haskins Labs, Conn. USA SR-86/87, 1-24.

Repp, B.H., and Williams, D.R. (1986). Categorical tendencies in imitating self-produced isolated vowels. *Status Report on Speech Research.* Haskins Labs, Conn. USA, SR-86/87, 55-70.

Riesz, R.R. (1928). Differential intensity sensitivity of the ear for pure tones. *Phys. Rev.* **31**, 867-875.

Robinson, D.E., and Watson, C.E. (1972). Psychophysical methods in modern psychoacoustics. In *Foundations of Modern Auditory Theory,* Vol. 2 (ed. J.V. Tobias). Academic Press, New York.

Robinson, D.W., and Dadson, R.S. (1956). A redetermination of the equal-loudness relations for pure tones. *Brit. J. Applied Physics.* **7**, 166-181.

Rose, J.E., Brugge, J.F., Anderson, D.J., and Hind, J.E. (1967). Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *J. Neurophysiol.* **30**, 769-793.

Rosenberg, A.E. (1971). Effect of glottal pulse shape on the quality of natural vowels. *JASA.* **49**, 583-590.

Rosenthal, R., and Rosnow, R.L. (1985). *Contrast Analysis: Focused Comparisons in the Analysis of Variance.* Cambridge University Press, Cambridge.

Rouiller, E., de Ribaupierre, Y., Morel, A., and de Ribaupierre, F. (1983). Intensity functions of single unit responses to tone in the medial geniculate body of cat. *Hear. Res.* **11**, 235-247.

Rye, J.M., and Holmes, J.N. (1982). A versatile software parallel-formant speech synthesizer. *JSRU Research Report* No. 1016.

Sachs, M.B., and Abbas, P.J. (1974). Rate versus level functions for auditory-nerve fibres in cats: tone-burst stimuli. *JASA.* **56**, 1835-1847.

Scharf, B. (1970). Critical Bands. In *Foundations of Modern Auditory Theory*, Vol. 1 (ed. J.V. Tobias). Academic Press, New York.

Sekey, A., and Hanson, B.A. (1984). Improved 1-Bark bandwidth auditory filter. *JASA.* **75**, 1902-1904. [L].

Sergeant, R.L., and Harris, J.D. (1962). Sensitivity to unidirectional frequency modulation. *JASA.* **34**, 1625-1628.

Shower, E.G., and Biddulph, R. (1931). Differential pitch sensitivity of the ear. *JASA.* **2**, 275-287.

Spenner, H., and Urbas, J.V. (1986). Forward masking and the perception of stop consonants: psychophysical and neurophysiological experiments. *Proc. IOA Autumn Conf.* **8**, 37-44.

Sreenivas, T.V., and Rao, P.V.S. (1982). Analysis of non-stationary voiced segments in speech signals. In *The Representation of Speech in the Peripheral Auditory System*, (eds, R. Carlson and B. Granström). Elsevier Biomedical Press, Amsterdam.

Stephens, S.D.G. (1973a). Auditory temporal integration as a function of intensity. *J. Sound Vib.* **30**, 109-126.

Stephens, S.D.G. (1973b). Some experiments on the detection of short duration stimuli. *Brit. J. Audiol.* **7**, 81-94.

Stevens, K.N., and Klatt, D.H. (1971). The role of formant transitions in the voice-voiceless distinction for stops. *QPR RLE MIT.* **101**, 188-197.

Tartter, V.C., and Eimas, P.D. (1975). The role of auditory feature detectors in the perception of speech. *Percept. Psychophys.* **18**, 293-298.

Tsumura, T., Sone, T., and Nimura, T. (1973). Auditory detection of a frequency transition. *JASA.* **53**, 17-25.

Viemeister, N.F. (1972). Intensity discrimination of pulsed sinusoids: the effects of filtered noise. *JASA.* **51**, 1265-1269.

Viemeister, N.F. (1974). Intensity discrimination of noise in the presence of band-reject noise. *JASA.* **56**, 1594-1600.

Wang, W.S.Y. (1959). Transition and release as perceptual cues for final plosives. *J. Speech Hear. Res.* **2**, 66-73.

Watson, C.S., and Gengel, R.W. (1969). Signal duration and signal frequency in relation to auditory sensitivity. *JASA.* **46**, 989-997.

Wells, J.C. (1987). *Computer-coded phonetic transcription.* A report on the Phonetic Transcription discussions held as part of the Workshop on Labelling, Transcription and Management Methods for Speech Databases held at University College, London.

Wier, C.C., Jesteadt, W., and Green, D.M. (1977). Frequency discrimination as a function of frequency and sensation level. *JASA.* **61**, 178-184.

Wilson, J.P. (1983). Application for a project grant to RSRE, Malvern for work on *Psychoacoustical constraints on automatic speech recognition.*

Winslow, R.L., Barta, P.E., and Sachs, M.B. (1987). Rate coding in the auditory-nerve. In *Auditory Processing of Complex Sounds*, (eds W.A. Yost and C.S. Watson). Erlbaum, Hillsdale, New Jersey.

Zwicker, E. (1956). Die elementaren Grundlagen zur Bestimmung der Informationkapazität des Gehörs. *Acustica.* **6**, 365-381.

Zwicker, E. (1970). Masking and psychological excitation as consequences of the ear's frequency analysis. In *Frequency Analysis and Periodicity Detection in Hearing*, (eds R. Plomp and G.F. Smoorenburg). Sijthoff, Leiden.

Zwicker, E., and Schorn, K. (1978). Psychoacoustical tuning curves in audiology. *Audiology.* 17, 120-140.